QSAR Models for Nitrogen Containing Monophosphonate and Bisphosphonate Derivatives as Human Farnesyl Pyrophosphate Synthase Inhibitors Based on Monte Carlo Method

Authors
Parvin Kumar¹, Ashwani Kumar², Jayant Sindhu³, Sohan Lal¹

Affiliations
1 Department of Chemistry, Kurukshetra University, Kurukshetra, Haryana, India
2 Department of Pharmaceutical Sciences, Guru Jambheshwar University of Science and Technology, Hisar, Haryana, India
3 K.M. Govt. College, Narwana, Haryana, India

Key words
Human farnesyl pyrophosphate synthase, Phosphonates, QSAR, IIC, CORAL, SMILES

Introduction
For the drug development, farnesyl pyrophosphate synthase (FPPS) enzyme has been established as a molecular target [1]. FPPS (EC 2.5.1.10), a member of E-family of the prenyltransferases, is an important enzyme in the mevalonate pathway, the elite method of isoprenoid synthesis in animals, concerned with cholesterol biosynthesis and post-translational modification of signaling proteins [2–4]. The human farnesyl pyrophosphate synthase (hFPPS) catalyzes the biosynthesis of the C-15 isoprenoid farnesyl pyrophosphate. QSAR studies by using Monte Carlo method for human farnesyl pyrophosphate synthase inhibitors has been carried out using balance of correlation technique with Index of ideality correlation. For construction of QSAR models, six random splits were prepared from the data of 73 phosphonates and hybrid optimal descriptors procured from graph (HFG) and SMILES based notations were employed. The developed QSAR models have robustness, good fitting ability, generalizability and internal predictive ability. The external predictive ability has been certified by testing various precedents. The values of $R^2$, IIC, $Q^2$ and $\Delta R^2_{int}$ for the best model are 0.9304, 0.9614, 0.9061 and 0.0861 respectively. The developed QSAR models met with the specified standards given in OECD guideline and applicability domain. The structural feature promoters for the end point increase and promoters for end point decrease have been extracted. The predicted $pIC_{50}$ for the new proposed compounds have also been reported.

Supporting Information
This article is available online at http://www.thieme-connect.de/products

ABSTRACT
Human farnesyl pyrophosphate synthase (hFPPS) is a well-settled therapeutic target and it is an enzyme of the mevalonate pathway which catalyzes the biosynthesis of the C-15 isoprenoid farnesyl pyrophosphate. QSAR studies by using Monte Carlo method for human farnesyl pyrophosphate synthase inhibitors has been carried out using balance of correlation technique with Index of ideality correlation. For construction of QSAR models, six random splits were prepared from the data of 73 phosphonates and hybrid optimal descriptors procured from graph (HFG) and SMILES based notations were employed. The developed QSAR models have robustness, good fitting ability, generalizability and internal predictive ability. The external predictive ability has been certified by testing various precedents. The values of $R^2$, IIC, $Q^2$ and $\Delta R^2_{int}$ for the best model are 0.9304, 0.9614, 0.9061 and 0.0861 respectively. The developed QSAR models met with the specified standards given in OECD guideline and applicability domain. The structural feature promoters for the end point increase and promoters for end point decrease have been extracted. The predicted $pIC_{50}$ for the new proposed compounds have also been reported.

Supporting Information
This article is available online at http://www.thieme-connect.de/products

ABSTRACT
Human farnesyl pyrophosphate synthase (hFPPS) is a well-settled therapeutic target and it is an enzyme of the mevalonate pathway which catalyzes the biosynthesis of the C-15 isoprenoid farnesyl pyrophosphate. QSAR studies by using Monte Carlo method for human farnesyl pyrophosphate synthase inhibitors has been carried out using balance of correlation technique with Index of ideality correlation. For construction of QSAR models, six random splits were prepared from the data of 73 phosphonates and hybrid optimal descriptors procured from graph (HFG) and SMILES based notations were employed. The developed QSAR models have robustness, good fitting ability, generalizability and internal predictive ability. The external predictive ability has been certified by testing various precedents. The values of $R^2$, IIC, $Q^2$ and $\Delta R^2_{int}$ for the best model are 0.9304, 0.9614, 0.9061 and 0.0861 respectively. The developed QSAR models met with the specified standards given in OECD guideline and applicability domain. The structural feature promoters for the end point increase and promoters for end point decrease have been extracted. The predicted $pIC_{50}$ for the new proposed compounds have also been reported.

Supporting Information
This article is available online at http://www.thieme-connect.de/products

ABSTRACT
Human farnesyl pyrophosphate synthase (hFPPS) is a well-settled therapeutic target and it is an enzyme of the mevalonate pathway which catalyzes the biosynthesis of the C-15 isoprenoid farnesyl pyrophosphate. QSAR studies by using Monte Carlo method for human farnesyl pyrophosphate synthase inhibitors has been carried out using balance of correlation technique with Index of ideality correlation. For construction of QSAR models, six random splits were prepared from the data of 73 phosphonates and hybrid optimal descriptors procured from graph (HFG) and SMILES based notations were employed. The developed QSAR models have robustness, good fitting ability, generalizability and internal predictive ability. The external predictive ability has been certified by testing various precedents. The values of $R^2$, IIC, $Q^2$ and $\Delta R^2_{int}$ for the best model are 0.9304, 0.9614, 0.9061 and 0.0861 respectively. The developed QSAR models met with the specified standards given in OECD guideline and applicability domain. The structural feature promoters for the end point increase and promoters for end point decrease have been extracted. The predicted $pIC_{50}$ for the new proposed compounds have also been reported.
like alendronate, risedronate and pamidronate have been revealed to inhibit FPPS enzyme [9–11].

Recently, J. Park et al. [12] discovered the biological importance of the allosteric pocket, near the IPP substrate binding site of the hFPPS active site, and exhibited that FPP product can bind to this allosteric pocket. This binding sealed the enzyme in an inactive conformation and therefore, supplying a feedback mechanism for regulating the intra-cellular levels of isoprenoid biosynthesis in vivo. To understand this mechanism, the thienopyrimidine-based monophosphonate (ThP-MP) or N-containing monophosphonates (NMps) have been designed and reported by J. Park et al. [1, 3] to get the essential minimal pharmacophore that is required for allosteric inhibition of hFPPS.

Therefore, many scientists are engaged to find out the liaison between the structure and the inhibition activities of phosphonate inhibitors [13–16]. The most successful method for constructing this liaison is Quantitative Structure Activity Relationship (QSAR) method, which is used to predict the drug activities and information for designing new potential drugs [17–21]. In the present era of drug design, the significance of quantitative structure activity relationship method is well recognized in view of the fact that QSAR know how to build the early forecast of activity-related characteristics of drug molecules and can eradicate the molecules with obnoxious properties [22, 23]. Literature survey shows that 3D-QSAR studies have been reported between NBPs and FPPS [24–31]. However, the majority of QSAR study was performed in non-mammalian species although Fernández et al. [32] and Liu et al. [33] reported the 3D-QSAR of N-BPs against hFPPS. It should be more reliable to take into account thienopyrimidine-based monophosphonate (ThP-MP) and N-BPs to study the QSAR of phosphonate.

Recent published papers reveal that the simplified molecular input line entry system (SMILES) is a substitute to classical QSAR methods and it can be used for the prediction of molecular structures with appropriate end point or activity [34–49]. In all the QSAR models, depending on Monte Carlo optimization method, the pertinent activity is treated as random event [50–53]. In the light of these facts and in continuation of our work [48, 54–63] on biological important heterocyclic compounds, we herein report Monte Carlo method based QSAR studies of 73 compounds i.e. thienopyrimidine-based monophosphonate (ThP-MP) and N-BPs to study the QSAR of phosphonate.

The balance of correlations is a technique described in the literature [66–69]. The crux of this technique is building up of a model via the Monte Carlo optimization of the following target function (TF):

\[
TE = R_{\text{training}} + R_{\text{invisible-training}} - |R_{\text{training}} - R_{\text{invisible-training}}| \times \text{Const}
\]

The index of ideality of correlation (IIC) is calculated with the following formula:

\[
\text{IIC} = r_{\text{calibration}} \times \frac{\min(\text{MAE}_{\text{calibration}}^+ \times \text{MAE}_{\text{calibration}}^-)}{\max(\text{MAE}_{\text{calibration}}^+ \times \text{MAE}_{\text{calibration}}^-)}
\]

The \(r_{\text{calibration}}\) is the correlation coefficient value between experimental and calculated values of an endpoint for the calibration set. MAE is mean absolute error which can be determined with the following equation:

\[
\text{MAE}_{\text{calibration}}^- = \frac{1}{N} \sum_{k=1}^{N} |\Delta_k| \quad \Delta_k < 0,
\]

\(N\) is the number of \(\Delta_k < 0\).
\[ \text{MAE}_{\text{calibration}} = \frac{1}{N} \sum_{k=1}^{N} |\Delta_k| \quad \text{for} \quad \Delta_k \geq 0 , \] \hspace{1cm} (5) \\

\[ + \text{N is the number of } \Delta_k \geq 0 \]

The quality of prediction (\( \Delta_4 \)) for one substance from a set can be estimated as the following:

\[ \Delta_k = \text{observed}_k - \text{calculated}_k \]

The optimal descriptor

Descriptors procured from either SMILES or molecular graph can be applied to symbolize the molecular structure. Literature survey reveals that “hybrid” demonstration of the molecular structure, i.e., by SMILES along with molecular graph, can grant a better model demonstrated by higher statistical quality than the model which is predicted by only SMILES or molecular graph\(^{[48],[70]}\). The hybrid optimal descriptor DCW, adopted for generating QSAR models for the pIC50, was determined as per the following equation:

\[ \text{Hybrid} \quad DCW (T, N_{\text{epoch}}) = \text{SMILES} \quad DCW (T, N_{\text{epoch}}) + \text{Graph} \quad DCW (T, N_{\text{epoch}}) \]

Where \( T \) is threshold and \( N_{\text{epoch}} \) is number of epochs used in monte carlo method optimization. Threshold helps in exclusion of use of rare molecular features.

SMILES based descriptors were calculated using following equation:

\[ \text{SMILES} \quad DCW (T, N_{\text{epoch}}) = \sum \text{CW} (S_k) + \sum \text{CW} (SS_k) + \sum \text{CW} (SSS_k) + \text{CW} (BOND) + \text{CW} (NOSP) + \text{CW} (HALO) + \text{CW} (HARD) + \text{CW} (PAIR) \]

Where, \( S_k \), \( SS_k \) and \( SSS_k \) represents local smile attributes while \( \text{BOND}, \text{NOSP}, \text{HALO}, \text{HARD} \) and \( \text{PAIR} \) represent global SMILES attributes and exhibit presence of double, triple or stereochemical bond, presence of nitrogen, oxygen, sulphur and phosphorus, presence of halogen as well as pairing between these attributes\(^{[71]}\).

Optimal descriptors based on graph were calculated using following equation:

\[ \text{Graph} \quad DCW (T, N_{\text{epoch}}) = \sum \text{CW} (EC0_k) + \sum \text{CW} (EC1_k) + \sum \text{CW} (EC2_k) + \sum \text{CW} (EC3_k) + \sum \text{CW} (VS2_k) + \text{CW} (NNC_k) + \text{CW} (C3) + \text{CW} (C4) + \text{CW} (C5) + \text{CW} (C6) \]

Where, \( EC0_k, EC1_k, EC2_k \) are Morgan’s extended connectivity indices for \( k^{th} \) vertex respectively. \( VS2_k \) is the sum of vertex degrees which take place at topological distance 2 relative to \( k^{th} \) vertex. NNC is the nearest neighbor code. The C3, C4, C5 and C6 are descriptors for presence of three-member (C3), four-member (C4), five-member (C5) and six-member rings (C6) respectively. In this study, hydrogen filled graphs (HFG) have been used.

CORALSEA software was used to prepare the QSAR models. The best model was scrutinized using the procedure given in literature\(^{[70]}\). The most prognostic combination of threshold (T) and number of epochs (N\(_{\text{epoch}}\)) for the six splits was executed from values 1-10 for T and 1-50 for N\(_{\text{epoch}}\). The model was built using balance of correlation technique with index of ideality of correlation (IIC) of Monte Carlo method. The weight of dr was 0.1. Start step of the optimization was 0.5 * CW(SA). Precision of the optimization was 0.1 * CW(SA) and weight of IIC was 0.1. Here, CW(SA) is weight of structural attribute (SA) at the start.

Possessing numerical information on above CW, DCW(T, N\(_{\text{epoch}}\)) can be computed for molecules of training and test set. These data can be used for calculation of pIC50 according to following Eq. (9):

\[ \text{pIC}_{50} = C0 + C1 \times \text{Hybrid} \quad DCW (T, N_{\text{epoch}}) \]

Validation of QSAR model

The most important objective of any QSAR modeling is to establish a sturdy model competent to predict the idiosyncrasy of new molecules in an objective, reliable and precise manner\(^{[49],[72]}\). Three methods are cited in the literature for evaluation of sturdiness and reliability of developed model. These are: (a) internal validation or cross-validation using the training set compounds (b) external validation using the test set compounds (c) data randomization or Y-scrumbling.

Leave-one-out (LOO) cross validation technique was used to develop models as an internal validation. Cross-validated Q\(^2\) interpret the predictive ability of the model\(^{[73],[74]}\). Higher the value of Q\(^2\) means better model prediction. The cross-validated Q\(^2\) is stated as:

\[ Q^2 = 1 - \frac{\sum (Y_{\text{obs}} - Y_{\text{pred}})^2}{\sum (Y_{\text{obs}} - \bar{Y}_{\text{train}})^2} \]

Where, \( Y_{\text{obs}} \) is observed property of the training set compounds, \( Y_{\text{pred}} \) is LOO-predicted property of the training set compounds and \( \bar{Y}_{\text{train}} \) is mean observed property of the training set compounds. The predictive ability of model is considered as acceptable when Q\(^2\) is greater than 0.5.

Similar methodology can be applied for external validation. The predictive ability of a model is determined by calculating Q\(^2\)\(_{\text{ext}}\) which is defined as:

\[ Q^2_{\text{ext}} = 1 - \frac{\sum (Y_{\text{obs(test)}} - Y_{\text{pred(test)}})^2}{\sum (Y_{\text{obs(test)}} - \bar{Y}_{\text{train}})^2} \]
Where, $Y_{\text{obs(test)}}$ is the observed property of the test set compounds, $Y_{\text{pred(test)}}$ is the predicted property of the test set compounds and $Y_{\text{train}}$ is the mean observed property of the training set compounds. The value of $Q^2_{\text{ext}}$ for an acceptable model should be greater than 0.5 test [75].

Y-randomization test was performed to inspect the robustness of the model. A parameter $cR^2_p$ penalizes the model $R^2$ for a small difference between squared mean correlation coefficient ($R^2$) of randomized models and squared correlation coefficient ($R^2$) of the non-randomized model [75]. The parameter $cR^2_p$ is defined as:

$$cR^2_p = 1 - \sqrt{R^2 - R^2}$$

(12)

For an acceptable QSAR model, the value of $cR^2_p$ should be greater than 0.5.

**Compliance with Applicability domain (AD) and OECD principles**

Applicability domain (AD) is another important aspect of a built QSAR model. It gives the information related to the biological, structure and physiochemical properties to make the prophecy for new compounds. AD is especially significant because it is applied for the evaluation of the authenticity of the developed QSAR model. Compounds from the training set can be used to interpret the AD of the developed QSAR model. The predictions of a QSAR/QSPR model are more authentic when predicted molecules are within applicability. For defining AD, procedure reported in literature was applied [65].

To assist the reflection of a QSAR model for regulatory purposes [76], it should satisfy the five principal as: 1) a defined endpoint: hFPPS inhibitory activity as definite endpoint, 2) an unambiguous algorithm: Monte Carlo method as unambiguous algorithm, 3) a defined domain of applicability: percentage of molecular features with defined role as domain of applicability, 4) appropriate measures of goodness-of-fit, robustness and predictivity: High values of $R^2$, $Q^2$, $cR^2_p$, $R^2_{\text{av}}$ and $\Delta R^2_p$, as suitable measures of goodness-of-fit, robustness and predictivity 5) a mechanistic interpretation, if possible: List of molecular features responsible for increase and decrease of activity are applied for mechanistic interpretation.

**Result and Discussion**

Six random splits were generated from the data set presented in Table 15 (supplementary information). All sets were carefully prepared so the ranges are approximately equivalent for each sub-set. The hybrid optimal descriptor DCW was adopted for generating QSAR models. The outcomes of the applied methodology for defining AD reveal that maximum numbers of molecules are within the defined AD. So, all the studied compounds had typical behavior and all were incorporated in developing QSAR models.

The best QSAR models for six different splits with other statistical parameters of all six equations are given in Table 1.

The statistical parameters given in Table 1 clearly indicate that the values of statistical criteria of all six equations are significant for each subset i.e. training, invisible-training, calibration and validation sets. The statistical trait of calibration set ($R^2 = 0.9304$ and $Q^2 = 0.9061$) for the equation of split 4 was distinguished best, therefore the QSAR model expressed by this was judged to be the preeminent model. The good fitting ability and good internal predictive power of the described model has been implied by the $R^2$ value (0.7380) for the training sets and $Q^2$ value of calibration set. The value of $R^2$ is 0.7602 for validation sets which shows the excellent external predictive ability of the developed QSAR model. The authenticity and robustness of the QSAR models are justified by less difference between $R^2$ and $Q^2$ values.

Y-randomization process was applied to check the chance correlation, in which Y values were scrambled in 1000 trials in ten separate runs (Table 35, supplementary information). The resulting value of $cR^2_p$ was found more than 0.5 which authenticate that the developed models are free from chance correlation [75].

The built QSAR model was checked for the external predictive power by applying the benchmarks proposed by Golbraikh and Tropsha [74], Roy and Roy [77] and Ojha et al. [78]. The values of different proposed benchmarks are displayed in Table 45 (supplementary information) and it can be seen that these values are within specified ranges. Therefore, it can be said that these QSAR equations have good external predictive ability. Calculated activity and applicability domain of all compounds along with different sets are exhibited in supplementary Table 55 (supplementary information). The procedure described in literature was used for the calculation of applicability domain and it is based on split defect [66].

The plots between observed versus calculated activity and residuals versus observed activity are shown in Fig. 1.

**Mechanistic interpretation of developed QSAR models**

The correlation weights (CW) of structural features (SAk) give the information about the structural attributes which are responsible for the increase and decrease of the endpoint. If CW(SAk) is positive in all three probes then this feature will enhance the value of endpoint and if CW(SAk) is negative in all three probes then it will reduce the value of endpoint. Based on these considerations, some graph based structural attributes with stable positive values (pIC50 enhancer) for best QSAR model 3 are EC0-... three, EC1-... four, EC2-... one, NNC-... one, NNC-... four, O---... four, NNC-P---... four, EC2-C---... one, VS2-C---... six, etc. Similarly, some graph based structural attributes with stable negative values (pIC50 decreasing) for this model are EC0-C---... two, EC1-C---... four, EC0-C---... three, ECO-N---... two, EC1-C---... six, NNC-C---... 220, VS2-C---... nine, etc.

In the same manner, some pIC50 enhancer SMILES based attributes are(...O...(branched oxygen with branching), +++++N---B2 == (nitrogen with double bond), +++++N---O== (nitrogen with oxygen), +++++N---P==-(nitrogen with phosphorous), +++++O---B2 == (oxygen with double bond), +++++O---P==-(oxygen with phosphorous), +++++P---B2 == (phosphorous with double bond), +++++P---O==-(oxygen with phosphorous), +++++P---B2 == (phosphorous with double bond), ---(methyl group), O...(oxygen with double branching), O...(oxygen with branching), O...(oxygen with double bond and branching), P...(phosphorous with branching), P...(phosphorous), C4....O...(absence of four membered ring), C3....O...(absence of three membered ring), BOND10000000(one
It has been reported in literature [1, 3, 7–12] that the interactions as hFPPS inhibitors are more potent inhibitors than monophosphonates. So, we can say that the developed QSAR models in the present manuscript have better statistical quality as compared to the reported model in terms of internal as well as external prediction criteria. Also, the resent QSAR models are one parameter 2D QSAR models which are simple in interpretation and simple to apply.

The result of the reported QSAR models predicted the activity of reported bisphosphonates and monophosphonates accurately. Two 3D QSAR model developed by comparative molecular field analysis (CoMFA) method for the some part of our dataset has been described by Fernández et al. [32] and Liu et al.[33]. The values for statistical features used by Fernández et al. [32] to examine the model were n (training) = 42, R² = 0.967, Q² = 0.763, F = 63 and SEE = 0.11 and by Liu et al. [33] to examine the model were n (training) = 54, R² = 0.982, Q² = 0.849, F = 104 and SEE = 0.09. In the reported QSAR models, monophosphonates were not used, but have used to develop a relationship between monophosphonate and bisphosphonates conditioned by the protonation state of the functional groups but the method used in this work does not consider this aspect. It is a limitation of this work. Despite that the proposed QSAR models predicted the activity of reported bisphosphonates and monophosphonates accurately.
Conclusion

Monte carlo optimization method using CORAL software has been used successfully for designing a statistically robust QSAR models for human farnesyl pyrophosphate synthase inhibitors. The best QSAR model described by DCW(7,13) optimal descriptor was obtained from split 4. In the mechanistic interpretation, the presence of nitrogen with double bond, a phosphate group and oxygen with branched double bond were found promoter of endpoint. The dis-

Fig. 1 The plots between observed pIC50 versus calculated pIC50 and residuals versus observed pIC50.

Fig. 2 Graphical representation of calculated pIC50 from all models for all proposed compound.
cussed QSAR models satisfy the OECD conditions for a good QSAR model. The outcomes of the applied methodology for defining AD reveal that maximum numbers of molecules are within the defined AD. Thus, this approach can be used for generation of new potential human farnesyl pyrophosphate synthase inhibitors.

Acknowledgements
The authors are grateful to Dr Andrey A. Toropov and Dr Alla P. Toropova for providing CORAL software.

Conflict of Interest
The authors declare no conflicts of interest.

References


[35] Stoïciov V, Stoianović D, Tasić I et al. QSAR study of 2,4-dihydro-3H-1,2,4-triazol-3-ones derivatives as angiotensin II (1) receptor antagonists based on the Monte Carlo method. Struct Chem 2018; 29: 441–449


[55] Kumar P. A novel, facile, simple and convenient oxidative aromatization of Hantzsch 1, 4-dihydropyridines to pyridines using polymeric iodobenzene with KBr. Journal of Heterocyclic Chemistry 2010; 47: 1429–1433


[69] Toropova AP, Toropov AA. CORAL: Monte Carlo Method to Predict Endpoints for Medical Chemistry. Mini Reviews in Medicinal Chemistry 2018; 18: 382–391


