INTERNATIONAL JOURNAL OF LATEST TECHNOLOGY IN ENGINEERING, MANAGEMENT & APPLIED SCIENCE (IJLTEMAS)

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XIV, Issue XII, December 2025

Centrality Measures in QSPR Modelling of Antiviral Compounds for COVID-19

Pavithra M and Veena Mathad

Department of Studies in Mathematics, University of Mysore, Manasagangotri, Mysuru, India

DOI : https://doi.org/10.51583/IJLTEMAS.2025.1412000101

Received: 28 December 2025; Accepted: 02 January 2026; Published: 09 January 2026

ABSTRACT:

The Severe Acute Respiratory Syndrome Coronavirus-2 (SARS-CoV-2), causing COVID-19, lacks specific antiviral treatments, escalating the global health crisis. This study employs Quantitative Structure-Property Relationship (QSPR) modelling to explore eight physicochemical properties of antiviral compounds, including Arbidol, Chloroquine, Hydroxychloroquine, Lopinavir, Remdesivir, Ritonavir, Thalidomide, and Theaflavin. Centrality measures, used as molecular descriptors, quantify the relationship between molecular structure and physicochemical attributes in QSPR studies.

To address missing data for Remdesivir's Boiling Point (BP), Enthalpy of Vaporisation (E), and Flash Point (FP), correlation-based linear regression imputation was applied using descriptors like Normalised Harmonic Centrality Weight (r =0.976 for BP, 0.973 for E) and Eccentricity Weight (r =0.957 for FP), ensuring dataset integrity. Nine graph-based centrality measures were evaluated for their correlation with the physicochemical properties of these drugs. Pearson correlation analysis revealed strong positive correlations, notably Normalised Harmonic Centrality Weight with BP (0.978), E (0.974), and Polar Surface Area (PSA) (0.894), and Eccentricity Weight with Flash Point (0.959), Molar Refractivity (MR) (0.975), and Molar Volume (MV) (0.923). Conversely, Total Closeness Centrality Weight and Leverage Centrality Weight showed significant negative correlations (below -0.5). Single-predictor linear regression models were developed, with robustness assessed via predictive R^2 using leave-one-out cross-validation and the PRESS statistic. These models offer interpretable predictions of structural influences on physicochemical behaviour, aiding pharmaceutical researchers in predicting antiviral drug properties for COVID-19 before experimental validation.

Keywords: Centrality Measures, QSPR Modelling, Antiviral Drugs, Physicochemical Properties, Molecular Descriptors, Predictive Modelling

Mathematics Subject Classification AMS (2000): 05C12, 05C38.

INTRODUCTION

The global outbreak of COVID-19, caused by Severe Acute Respiratory Syndrome Coronavirus-2 (SARS-CoV-2), originated in Wuhan, China, on 31st December 2019 [1], and was classified as a pandemic by the World Health Organisation on 11th March 2020 [2]. The lack of specific antiviral agents poses a critical barrier to effective management of the disease, necessitating rapid identification of viable therapeutic options [3]. Antiviral compounds previously developed for infections such as SARS, MERS, and Influenza are currently under evaluation for their potential against SARS-CoV-2 [4, 5]. These include Arbidol [6], Chloroquine [7], Hydroxychloroquine [5], Lopinavir [8], Ritonavir [9], Thalidomide [10], Theaflavin [11], and Remdesivir [12], selected based on their pharmacological properties and preliminary indication of efficacy.

The foundation for selecting these compounds is rooted in their established mechanisms. Lopinavir and Ritonavir inhibit coronavirus proteases (papain-like and 3C-like), with clinical data suggesting reduced mortality in severe cases [13]. Arbidol, employed in Russia and China, targets Influenza A, Influenza B, and Hepatitis C, prompting its investigation despite limited global approval for COVID-19 [9]. Thalidomide's immunomodulatory effects make it suitable for addressing inflammatory complications [10]. Chloroquine and Hydroxychloroquine,

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XIV, Issue XII, December 2025

traditionally antimalarial drugs, exhibit antiviral and immune-modulating activities [7]. Theaflavin, derived from tea, shows early antiviral potential [11]. Remdesivir, originally for Ebola, disrupts SARS-CoV-2 replication, with trials indicating therapeutic promise [12, 14]. A preliminary study combining Lopinavir/Ritonavir, Arbidol, and Shufeng Jiedu Capsule (a Chinese herbal remedy) reported clinical benefits in patients, supporting exploration of combination therapies [6, 15].

This investigation applies Quantitative Structure-Property Relationship (QSPR) modelling, employing distance-based and degree-based centrality measures derived from molecular graph theory to predict eight physicochemical properties of the selected compounds: Boiling Point (BP, mmHg), Enthalpy of Vaporisation (E, kJ/mol), Flash Point (FP, °C), Molar Refractivity (MR, cm^3), Polar Surface Area (PSA, $Å^2$), Packing Volume (P, cm^3), Molar Volume (MV, cm^3), and Molecular Weight (MW, g/mol) [16, 17]. These centrality measures quantify structural influences on molecular behaviour, enabling prediction of properties critical for drug design and reducing dependence on resource-intensive laboratory experiments [18, 19].

A primary obstacle was the unavailability of experimental data for Remdesivir's boiling point, enthalpy of vaporisation, and flash point [20]. Lacking a full dataset, making QSPR models is unreliable, as incomplete data can compromise model accuracy and reliability [21, 22]. This was resolved through correlation-based linear regression imputation, utilising strongly correlated molecular descriptors to maintain dataset integrity [23, 24].

To address missing data in this investigation of QSPR analysis, we employed regression-based data substitution, achieving a correlation fit of 97%. This approach is supported by studies demonstrating its efficacy in predictive modelling [25] and its suitability for datasets with strong variable correlations [26]. While applicable to longitudinal data [27], its principles extend to QSPR's interdependent molecular descriptors. Compared to modern techniques [28], regression imputation remains a robust choice for our high-correlation context. The methodology integrates statistical analysis via Minitab and advanced data processing through MATLAB, structured in three phases: imputation of missing Remdesivir data, construction of QSPR models using single-predictor linear regression based on centrality measures, and validation of model accuracy via the Predicted Residual Error Sum of Squares (PRESS) statistic [29, 30], from leave-one-out cross-validation. This approach ensures robust and generalizable models [31-33].

Complex networks, consisting of nodes and links, are dominant in various fields like physics, biology, and social sciences; these include examples such as the internet, social media, and biological networks like protein-protein interaction links. Studying these networks is a crucial area of multidisciplinary research, as identifying the most influential nodes is theoretically and practically significant for understanding how information spreads; centrality is a key concept in network analysis used to determine a node's importance, and these measures are among the most common analytical techniques for identifying powerful nodes.

Several centrality measures have been proposed to evaluate the influence of nodes in networks, including degree, betweenness, closeness, leverage, harmonic, and eigenvector centrality. Degree Centrality ranks nodes by the number of connections they have, with those having more connections being considered more influential (1)[34]; a related measures, the Total Degree Centrality Weight ($DW(G)$), and Degree Centrality Weight ($DCW(G)$), measure the overall heterogeneity of the network's degree distribution, reflecting how evenly connections are distributed across nodes (11, 12). Closeness centrality measures how quickly a node can spread information to other nodes in the network (2)[35], while the Total Closeness Centrality Weight ($TCW(G)$) and Closeness Centrality Weight $\left(CW(G)\right)$ are global measures that characterise the network's overall compactness and communication efficiency (4, 5). Leverage Centrality considers a node's degree relative to its neighbours, operating on the principle that a node is central, if its neighbours depend on it for information (6) [36, 37]; the Leverage Centrality Weight ($LW(G)$) serves as a collective measure of the network's structural dependency (13). Harmonic Centrality calculates a node's importance by summating the inverse of its geodesic distances to all other nodes (7)[38]; the corresponding Harmonic Centrality Weight $\left(HW(G)\right)$ and Normalised Harmonic Centrality Weight $\left(HCW(G)\right)$ quantify the network's global influence potential and overall efficiency (14, 15). Finally, Eccentricity Centrality identifies the "weakest links" in a network, which can be targeted for either disruption or strengthening to improve flexibility (10)[39]; the Eccentricity Weight ($EW(G)$) provides insight

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XIV, Issue XII, December 2025

into the longest paths within the network (16), while the Eccentricity Centrality Weight ($ECW(G)$) quantifies the network's overall flexibility by summing the reciprocals of these path lengths (17).

Mathematical Preliminaries and Centrality Measures

Let $G = (V(G), E(G))$ be a simple, connected, finite, and undirected graph, where $V(G)$ is the set of vertices and $E(G)$ is the set of edges. The number of vertices and edges are denoted by $n = |V(G)|$ and $m = |E(G)|$, respectively. The degree of a vertex v, denoted by $d(v)$, is the number of edges incident to v. The distance between two vertices u and v, denoted by $d(u, v)$, is the length of the shortest path connecting them in G.

Centrality Measures

1.1.1 Degree Centrality ($D_C(v)$)[34]: Degree Centrality is a measure based on the number of connections a node has. The degree centrality of a vertex $v \in V(G)$ is defined as the number of vertices adjacent to v, which is nothing but the degree of v. This value is divided by the maximum possible degree of a vertex to normalise it. So, the Normalised Degree Centrality ($DC(v)$) of the vertex v is given by:

$$DC(v) = \frac{d(v)}{n-1} \quad \text{.... (1)}$$

1.1.2 Closeness Centrality ($C_C(u)$)[35]: Closeness Centrality is the reciprocal of the sum of distances from a vertex to all other vertices. For a vertex $u \in V(G)$ it is defined as:

$$C_C(u) = \frac{1}{\sum_{v \in V(G), v \neq u} d(u,v)} \quad \text{.... (2)}$$

This value is multiplied by the maximum possible degree of a vertex to normalise it. So, the normalised Closeness Centrality ($C_G(u)$) of a vertex u is given by:

$$C_G(u) = \frac{n-1}{\sum_{v \in V(G), v \neq u} d(u,v)} \quad \text{.... (3)}$$

The Total Closeness Centrality Weight ($TCW(G)$) of G is:

$$TCW(G) = \sum_{k=1}^{n} C_C(v_k) \quad \text{.... (4)}$$

The Closeness Centrality Weight ($CW(G)$) of G is:

$$CW(G) = \sum_{k=1}^{n} C_G(v_k) \quad \text{.... (5)}$$

1.1.3 Leverage Centrality ($L(v)$)[36, 37]: Leverage Centrality measures the relationship between a vertex's degree and the degrees of its neighbours. For a vertex $v \in V(G)$, it is defined as:

$$L(v) = \frac{1}{d(v)} \sum_{u \in N_v} \frac{d(v) - d(u)}{d(v) + d(u)} \quad \text{.... (6)}$$

Where N_v is the set of neighbours of v.

1.1.4 Harmonic Centrality ($HC(u)$)[38]: Harmonic Centrality is the sum of the inverse of distances from a vertex to all other vertices. For a vertex $u \in V(G)$, it is given by:

$$HC(u) = \sum_{v \in V(G), v \neq u} \frac{1}{d(u,v)} \quad \text{.... (7)}$$

The Normalised Harmonic Centrality ($NHC(u)$) of a vertex u is:

$$NHC(u) = \frac{1}{n-1} \sum_{v \in V(G), v \neq u} \frac{1}{d(u,v)} \quad \text{.... (8)}$$

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XIV, Issue XII, December 2025

1.1.5 Eccentricity ($e(v)$)[39]: The eccentricity of a vertex v, denoted by $e(v)$, is the maximum distance from v to any other vertex in the graph:

$$e(v) = \max_{u \in V(G)} d(v,u) \quad \text{.... (9)}$$

Eccentricity Centrality ($EC(v)$) is the reciprocal of the eccentricity of a vertex. For a vertex $v \in V(G)$, it is defined as:

$$EC(v) = \frac{1}{e(v)} = \frac{1}{\max_{u \in V(G)} d(v,u)} \quad \text{.... (10)}$$

The Total Degree Centrality Weight ($DW(G)$) of G is a measure of the network's heterogeneity, defined as the sum degrees of all vertices, which is equal to $2m$.

$$DW(G) = 2m \quad \text{.... (11)}$$

The Degree Centrality Weight ($DCW(G)$) of G is:

$$DCW(G) = \sum_{k=1}^{n} DC(v_k) \quad \text{.... (12)}$$

The Leverage Centrality Weight ($LW(G)$) of G is defined as the sum of the leverage centralities of all vertices:

$$LW(G) = \sum_{k=1}^{n} L(v_k) \quad \text{.... (13)}$$

The Harmonic Centrality Weight ($HW(G)$) of G is:

$$HW(G) = \sum_{u \in V(G)} HC(u) \quad \text{.... (14)}$$

The Normalised Harmonic Centrality Weight ($HCW(G)$)of G is:

$$HCW(G) = \sum_{u \in V(G)} NHC(u) \quad \text{.... (15)}$$

The Eccentricity Weight ($EW(G)$) of G is the sum of the eccentricities of all vertices:

$$EW(G) = \sum_{v \in V(G)} e(v) \quad \text{.... (16)}$$

The Eccentricity Centrality Weight ($ECW(G)$) of G is the sum of the eccentricity centrality of all vertices:

$$ECW(G) = \sum_{v \in V(G)} EC(v) = \sum_{v \in V(G)} \frac{1}{e(v)} \text{.... (17)}$$

Research influences fundamental principles of complex network analysis to characterise chemical structures. By employing a suite of well-established centrality measures-including degree, closeness, leverage, harmonic, and eccentricity centrality, the structural importance of individual atoms and their relationships within a molecular graph can be quantified. This detailed network-based characterisation using molecular graphs provides a powerful set of descriptors for QSPR analysis (Table 1). The present work outlines a robust methodology that integrates these centrality measure descriptors with statistical models. This integration ultimately paves the way for a deeper understanding of molecular properties and their direct application in designing novel drugs, with a particular focus on the urgent need for targeted rehabilitations for diseases like COVID-19.

MATERIALS AND METHODS

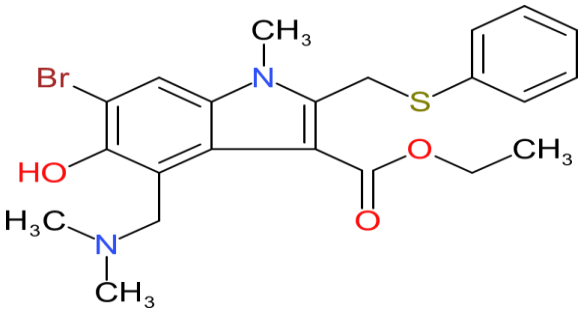
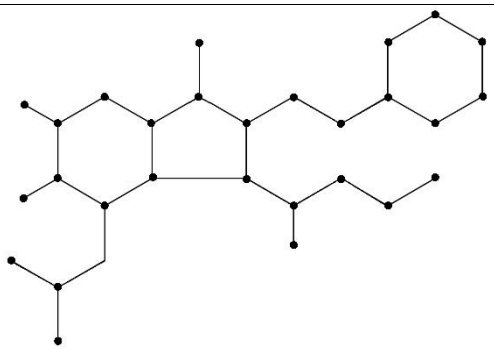
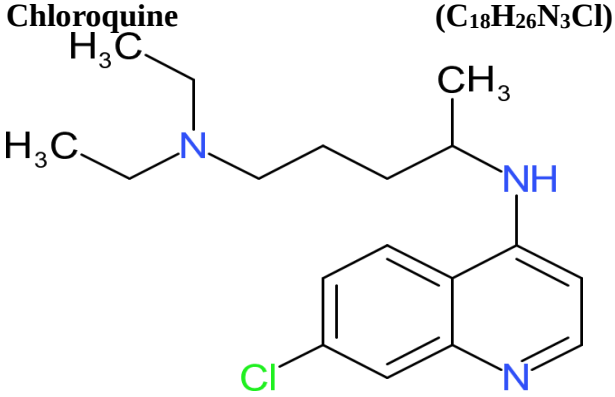
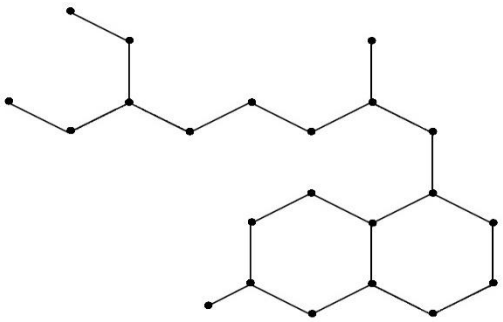
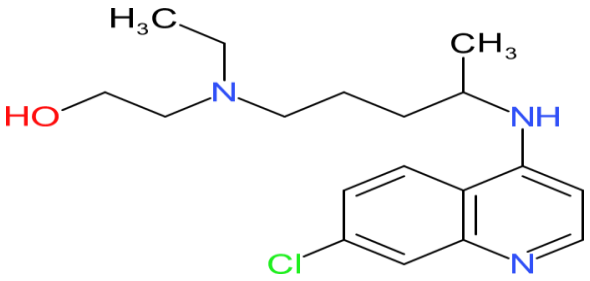
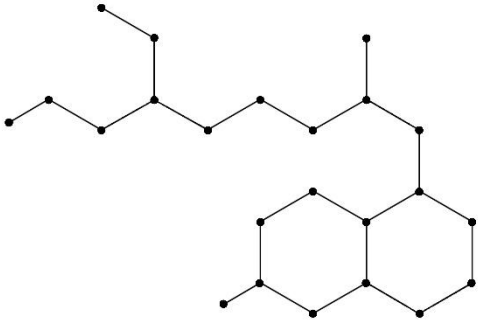
Data Collection and Preprocessing

The dataset for this study is taken from Kara et al. [20], which comprised eight diverse drug compounds: Arbidol, Chloroquine, Hydroxychloroquine, Lopanavir, Remdesivir, Ritonavir, Thalidomide, and Theaflavin. For each

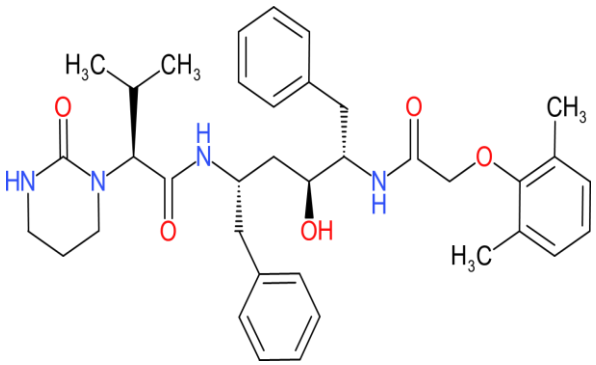
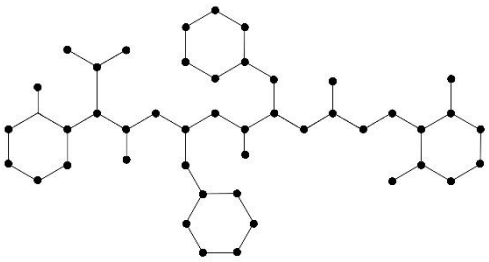
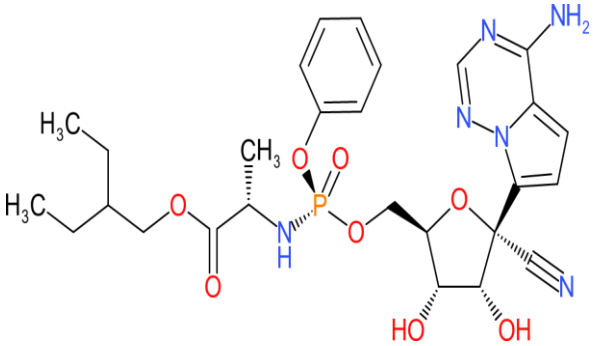
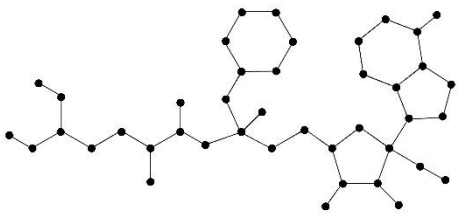
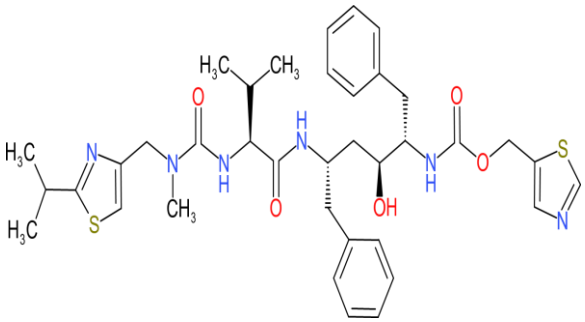
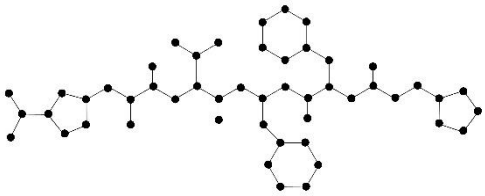
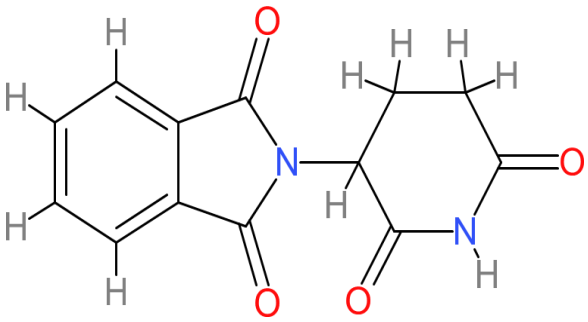
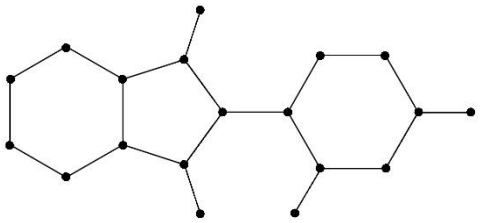
ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XIV, Issue XII, December 2025

compound, eight key physicochemical properties were considered: Boiling Point (BP, mmHg), Enthalpy of Vaporisation (E, kJ/mol), Flash Point (FP, °C), Molar Refractivity (MR, cm³), Polar Surface Area (PSA, Å²), Packing Volume (P, cm³), Molar Volume (MV, cm³), and Molecular Weight (MW, g/mol). Nine molecular graph-based centrality measures were used as descriptors: Total Degree Centrality Weight $(DW(G))$, Degree Centrality Weight $(DCW(G))$, Total Closeness Centrality Weight $(TCW(G))$, Closeness Centrality Weight $(CW(G))$, Leverage Centrality Weight $(LW(G))$, Harmonic Centrality Weight $(HW(G))$, Normalised Harmonic Centrality Weight $(HCW(G))$, Eccentricity Weight $(EW(G))$, Eccentricity Centrality Weight $(ECW(G))$. Table 2 and Table 3 contain the reference property data and the experimental and estimated physicochemical properties for the drugs.

Table 1: Antiviral Compounds with Corresponding Molecular and Graph Representations for QSPR Analysis

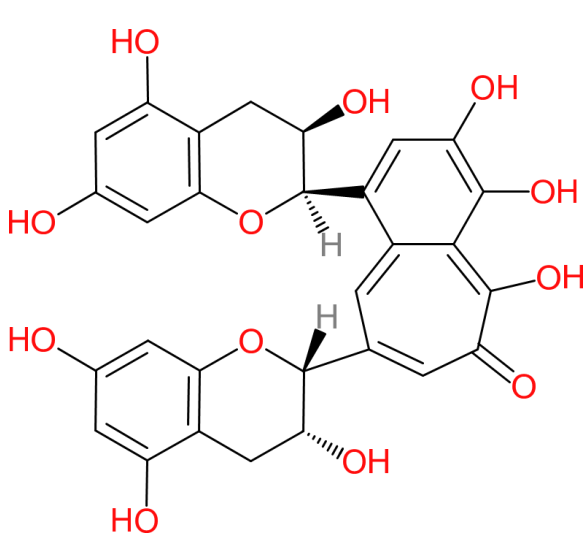
Molecular Structure	Molecular Graph Representations
Arbidol ($C_{22}H_{25}O_3N_2SBr$) 	
Chloroquine ($C_{18}H_{26}N_3Cl$) 	
Hydroxychloroquine ($C_{18}H_{26}N_3ClO$) 	

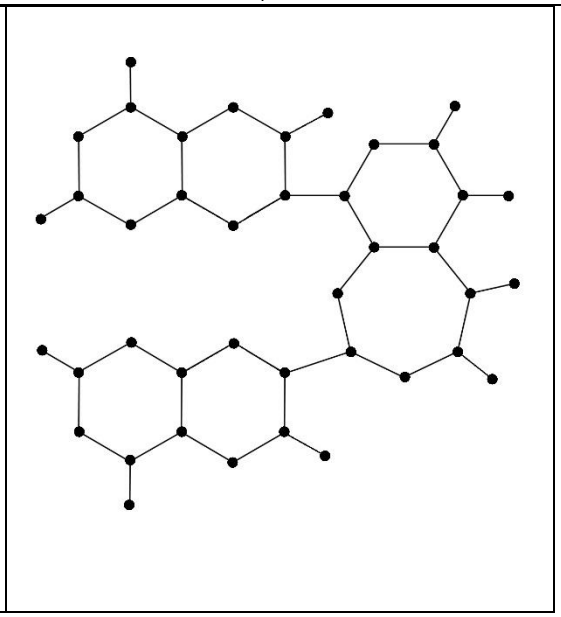
ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XIV, Issue XII, December 2025

Lopinavir ($C_{37}H_{48}O_5N_4$) 	
Remdesivir ($C_{27}H_{35}O_8N_6P$) 	
Ritonavir ($C_{37}H_{48}N_6O_5S_2$) 	
Thalidomide ($C_{13}H_{10}O_4N_2$) 	

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XIV, Issue XII, December 2025

Theaflavin ($C_{29}H_{24}O_{12}$)





Heatmaps were generated to visualise Standardised molecular descriptors and physicochemical properties of eight antiviral drugs (Figure 1 & Figure 2). The colour heatmaps use a blue-to-yellow gradient to highlight value differences. High descriptor values, such as elevated $DW(G)$ and $EW(G)$ for Ritonavir and Lopinavir, indicate greater molecular complexity, while high property values, like MW and PSA for Ritonavir, suggest increased size or polarity. These visualisations reveal structural and physicochemical patterns critical for QSPR modelling, assisting in the prediction of drug activity and chemo-kinetic behaviour.

To mitigate the impact of disparate scales observed across the raw dataset, continuous predictor variables were standardised for the linear regression modelling process. This was achieved by using the subtracted mean and then dividing by the standard deviation for continuous predictors. This process ensured that centrality measures contributed equitably to the determination of model coefficients, preventing undue bias from variables with larger absolute magnitudes. The physicochemical properties, serving as response variables, were retained in their original units for direct interpretability of predictions.

Table 2: Centrality Measures for Molecular Structure Representation in QSPR Modelling of Drug Compounds

Drugs	$DW(G)$	$DCW(G)$	$TCW(G)$	$CW(G)$	$HW(G)$	$HCW(G)$	$LW(G)$	$EW(G)$	$ECW(G)$
Arbidol	62.000	2.213	0.209	5.883	233.574	8.340	-2.865	274.000	3.165
Chloroquine	46.000	2.190	0.207	4.367	143.501	6.833	-1.399	224.000	2.234
Hydroxy Chloroquine	48.000	2.166	0.198	4.356	153.523	6.979	-2.198	250.000	2.188
Lopanavir	98.000	2.177	0.135	6.073	427.476	9.509	-3.635	653.000	3.334
Remdesivir	88.000	2.200	0.144	5.790	369.316	9.238	-3.623	574.000	3.013
Ritonavir	106.000	2.163	0.121	5.954	468.286	9.556	-4.467	844.000	3.058

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XIV, Issue XII, December 2025

Thalidomide	42.000	2.333	0.289	5.210	127.795	7.099	-1.499	135.000	2.763
Theaflavin	92.000	2.300	0.2070	6.707	403.023	10.077	-4.134	481.000	3.599

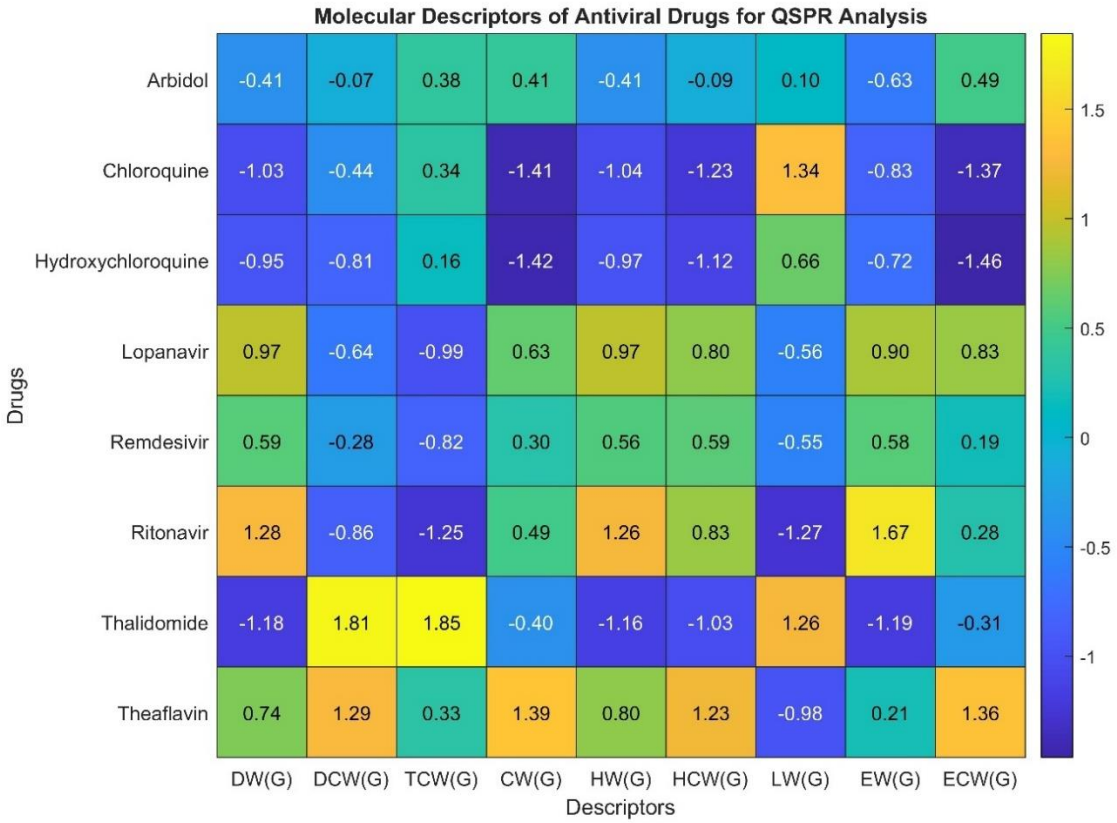


Figure 1: Molecular Descriptors of Antiviral Drugs for QSPR Analysis.

Table 3: Experimental and Estimated Physicochemical Properties of Drugs Used in QSPR Analysis

Drugs	Boiling Point	Enthalpy of Vaporisation	Flash Point	Molar Refractivity	Polar Surface Area	Packing Volume	Molar Volume	Molecular Weight
	BP (mmHg)	E (kJ/mol)	FP (ºC)	MR (cm^3)	PSA ($Å^2$)	P (cm^3)	MV (cm^3)	MW (g/mol)
Arbidol	591.80	91.50	311.7	121.90	80.00	48.30	347.30	477.40
Chloroquine	460.60	72.10	232.3	97.40	28.20	38.60	287.90	319.90
Hydroxy-Chloroquine	516.70	83.00	266.3	99.00	48.40	39.20	285.40	335.90
Lopinavir	924.20	140.80	512.7	179.20	120.00	71.00	540.50	628.80
Remdesivir	-	-	-	149.50	204.00	59.30	409.00	620.60

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XIV, Issue XII, December 2025

Ritonavir	947.00	144.40	526.6	198.90	202.00	78.90	581.70	720.90
Thalidomide	487.80	79.40	248.8	65.20	83.60	25.90	161.00	258.23
Theaflavin	1003.9	156.50	336.5	137.30	218.00	54.40	301.00	564.50

Note: BP: Boiling Point; E: Enthalpy of Vaporisation; FP: Flash Point; MR: Molar Refractivity; PSA: Polar Surface Area; P: Packing Volume; MV: Molar Volume; MW: Molecular Weight. Missing values were estimated via correlation-based regression.

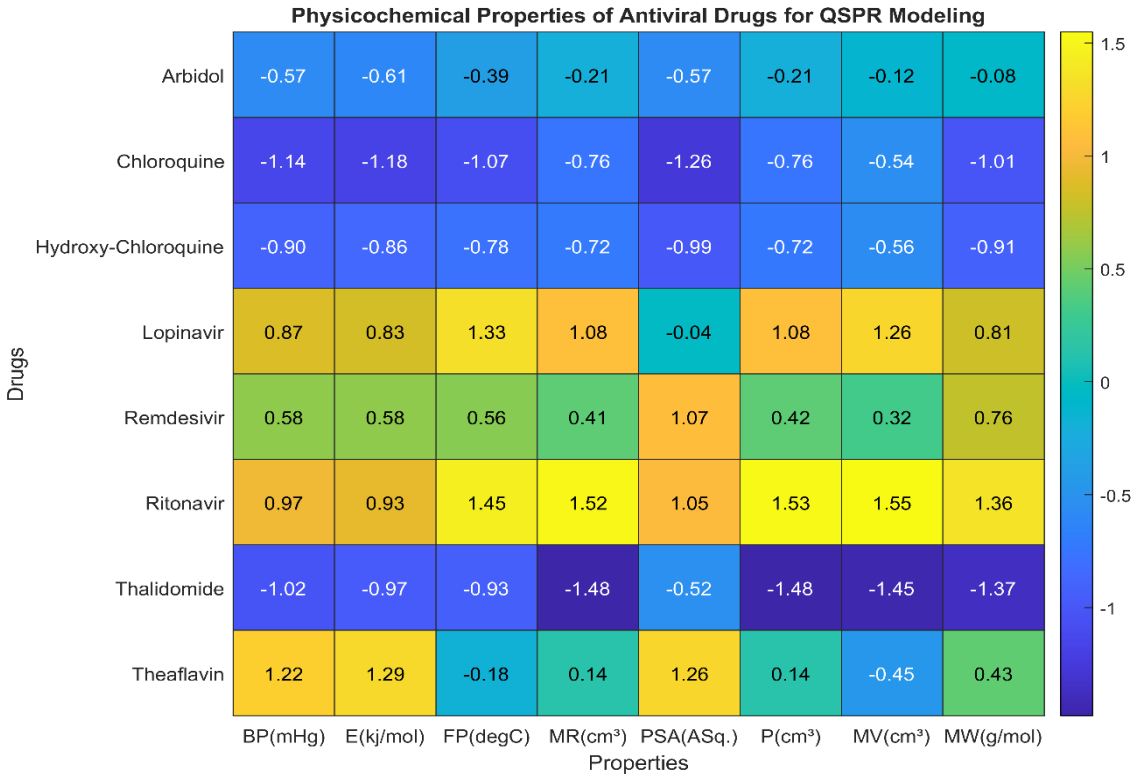


Figure 2: Physicochemical Properties of Antiviral Drugs for QSPR Modelling

Imputation of Missing Physicochemical Properties

Missing values for Boiling Point (BP), Enthalpy of Vaporisation (E), and Flash Point (FP) for remdesivir were imputed to complete the dataset for subsequent QSPR modelling. This imputation was performed using Pearson correlation-based simple linear regression models, the influence of strongest linear relationships observed between these properties and selected molecular descriptors. The selection of descriptors for imputation was based on the highest absolute Pearson correlation coefficients ($|r|$>0.95) obtained from preliminary correlation analyses (Table 4). Specifically, the $HCW(G)$ centrality measure was selected for BP (r =0.976) and E (r =0.973), while $EW(G)$ was chosen for FP (r = 0.957). Linear regression models were then fitted using these highly correlated predictors, Utilising data from other drugs in the dataset.

Table 4a: Pearson Correlation Coefficients for Properties Used in Remdesivir Data Imputation.

	$DW(G)$	$DCW(G)$	$TCW(G)$	$CW(G)$	$HW(G)$	$HCW(G)$	$LW(G)$	$EW(G)$
$DCW(G)$	-0.258							
$TCW(G)$	-0.804	0.769						

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XIV, Issue XII, December 2025

$CW(G)$	0.793	0.273	-0.33					
$HW(G)$	1	-0.234	-0.788	0.808				
$HCW(G)$	0.953	-0.03	-0.632	0.927	0.959			
$LW(G)$	-0.963	0.209	0.738	-0.828	-0.965	-0.959		
$EW(G)$	0.959	-0.436	-0.887	0.61	0.951	0.83	-0.897	
$ECW(G)$	0.749	0.303	-0.28	0.992	0.765	0.895	-0.772	0.551
BP	0.97	-0.056	-0.66	0.852	0.975	**0.976***	-0.954	0.874
E	0.958	-0.02	-0.63	0.855	0.964	**0.973***	-0.948	0.856
FP	0.917	-0.421	-0.85	0.608	0.909	0.782	-0.83	**0.957***
MR	0.949	-0.525	-0.92	0.621	0.941	0.831	-0.895	0.975
PSA	0.85	0.129	-0.499	0.817	0.856	0.894	-0.883	0.763
P	0.949	-0.524	-0.92	0.621	0.941	0.831	-0.895	0.975
MV	0.844	-0.673	-0.939	0.445	0.831	0.674	-0.76	0.923
MW	0.978	-0.375	-0.86	0.749	0.974	0.919	-0.957	0.958

Note: Values represent Pearson correlation coefficients between pairs of physicochemical properties and molecular descriptors deduced from Table 1 & Table 2. Bolded values indicate the highest correlation used for regression-based estimation of missing data.

Table 4b: Pearson Correlation Coefficients for Properties Used in Remdesivir Data Imputation (columns continued).

	$ECW(G)$	BP	E	FP	MR	PSA	P	MV
$DCW(G)$								
$TCW(G)$								
$CW(G)$								
$HW(G)$								
$HCW(G)$								
$LW(G)$								
$EW(G)$								
$ECW(G)$								
BP	0.817							

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XIV, Issue XII, December 2025

E	0.819	0.999						
FP	0.582	0.815	0.793					
MR	0.577	0.849	0.825	0.955				
PSA	0.746	0.915	0.925	0.649	0.675			
P	0.577	0.849	0.825	0.955	1	0.676		
MV	0.411	0.696	0.664	0.945	0.967	0.487	0.967	
MW	0.698	0.912	0.894	0.926	0.969	0.812	0.969	0.889

Note: Values represent Pearson correlation coefficients between pairs of physicochemical properties and molecular descriptors deduced from Table 1 & Table 2. Bolded values indicate the highest correlation used for regression-based estimation of missing data.

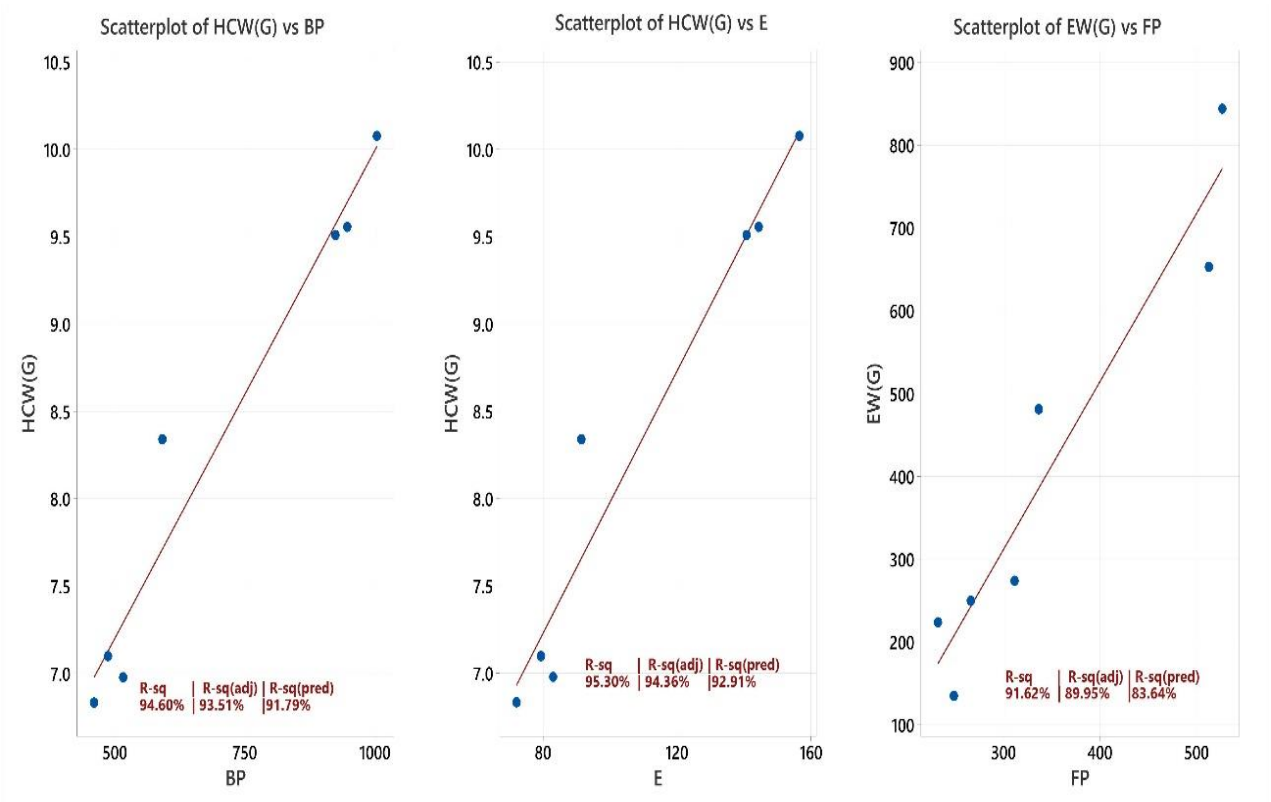


Figure 3: Regression Plots Illustrating Descriptor-Property Relationships Utilized for Remdesivir Data Imputation.

The Pearson correlation coefficients for the properties and descriptors used in this imputation are presented in Table 4. The resulting regression equations were subsequently applied to remdesivir's known $HCW(G)$ and $EW(G)$ values to estimate its missing BP, E, and FP values, ensuring robust estimations for subsequent quantitative structure-property relationship modelling. Table 5 summarises these imputed values. For a visual representation of the regression models employed for imputation, scatter plots illustrating these descriptor-property relationships are presented in Figure 3 contributing to robust fit accuracies & prediction ability.

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XIV, Issue XII, December 2025

Table 5: Imputation of Missing Physicochemical Properties for Remdesivir Using Correlation-Based Regression.

Property	Strongest Correlated Descriptor	Regression Equation	Input Descriptor Value	Estimated Property Value
Boiling Point (BP), in mmHg	$HCW(G)$	BP = -717 + 170.4 x $HCW(G)$	$HCW(G)$ =9.238375	857.22 mmHg
Enthalpy of Vaporisation (E), in kJ/mol	$HCW(G)$	E = -100.5 + 25.20 x $HCW(G)$	$HCW(G)$ =9.238375	132.31 kJ/mol
Flash Point (FP), in ºC	$EW(G)$	FP =163.4+ 0.4512 x $EW(G)$	$EW(G)$ = 574	422.39 ºC

Computational and Statistical Parameters

All statistical analysis and regression model developments were performed using statistical programmes, influencing its robust capabilities for linear regression. For each physicochemical property, potential QSPR models were developed by identifying the most statistically significant single centrality measure descriptor.

For all developed linear regression models, the predictive performance was rigorously assessed using the coefficient of determination R^2, adjusted R^2 (R^2_{adj}), and critically, the predictive R^2 $\left(R^2_{pred}\right)$. The R^2_{pred} was calculated using the PRESS (Predicted Residual Error Sum of Squares) statistic, which is derived from a leave-one-out cross-validation approach. In this method, each data point is successively removed from the dataset, and a model is built using the remaining n−1 data points to predict the removed observation. This process is repeated for all data points (includes eight observations for each of nine measures after data imputation), providing an unbiased estimate of the model's predictive ability for external data. A higher $\left(R^2_{pred}\right)$ indicates a more robust and generalizable model. The standard error of regression (S) and F-statistic with its associated p-value were also reported to evaluate model fit and overall statistical significance. The significance of individual predictor coefficients was assessed via p-values ($p < 0.05$). The resulting regression equations, derived from these models, were used to calculate the predicted values for the physicochemical properties, and these predictions were subsequently compared with the actual experimental properties to assess model accuracy and identify deviations.

Exploratory Data Analysis (EDA)

Exploratory Data Analysis was conducted to systematically explore the relationships between physicochemical properties and molecular descriptors, providing both statistical and visual insights crucial for informing subsequent model selection.

Table 6 presents the fundamental descriptive statistics, including mean (μ), standard deviation (σ), minimum, and maximum values, for all 17 centrality measures and physicochemical properties measured across the eight drug molecules, including remdesivir. This statistical summary illustrates the inherent variability and range of each descriptor and property within the dataset. Notably, significant scale disparities are evident, such as between $DW(G)$ (μ= 72.75, σ=26.01) and Boiling Point (BP: μ = 723.7, σ=230.4).

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XIV, Issue XII, December 2025

Table 6: Descriptive Statistics of Centrality Measures and Physicochemical Properties

Variable	Mean (μ_y)	Standard Deviation (σ_y)	Minimum (Min)	Maximum (Max)
$DW(G)$	72.750	26.0100	42.0000	106.0000
$DCW(G)$	2.2178	0.0638	2.1633	2.3330
$TCW(G)$	0.1892	0.0544	0.1215	0.2895
$CW(G)$	5.5430	0.8350	4.3560	6.7070
$HW(G)$	290.80	141.1000	127.8000	468.3000
$HCW(G)$	8.4540	1.3220	6.8330	10.0770
$LW(G)$	-2.978	1.1770	-4.4680	-1.3990
$EW(G)$	429.40	248.0000	135.0000	844.0000
$ECW(G)$	2.9190	0.5000	2.1880	3.5990
BP (mmHg)	723.70	230.4000	460.6000	1003.9000
E (kJ/mol)	112.50	34.2000	72.1000	156.5000
FP (ºC)	357.20	116.6000	232.3000	526.6000
MR (cm^3)	131.10	44.5000	65.2000	198.9000
PSA ($Å^2$)	123.00	75.4000	28.2000	218.0000
P(cm^3)	51.950	17.6600	25.9000	78.9000
MV (cm^3)	364.20	140.4000	161.0000	581.7000
MW(g/mol)	490.80	169.8000	258.2000	720.9000

Understanding the linear relationships between all variables was a foundational step. A comprehensive Pearson correlation analysis was performed on the complete dataset (post-missing data substitution) to elucidate these relationships. The Pearson correlation coefficient, ranging from -1 to 1, provides a quantitative measure of both the strength and direction of linear associations (Figure 5).

Initially, the intercorrelations among the eight physicochemical properties themselves were examined. Within the physicochemical properties themselves, Figure 4 illustrates a Pearson correlation matrix highlighting strong linear relationships (e.g., Boiling Point with Enthalpy of Vaporisation, Molar Refractivity with Packing Volume); while indicative of shared molecular influences, these intrinsic inter-property relationships also contribute to collinearity, posing a significant consideration for direct multi-linear regression (MLR) modelling, especially with the limited available dataset.

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XIV, Issue XII, December 2025

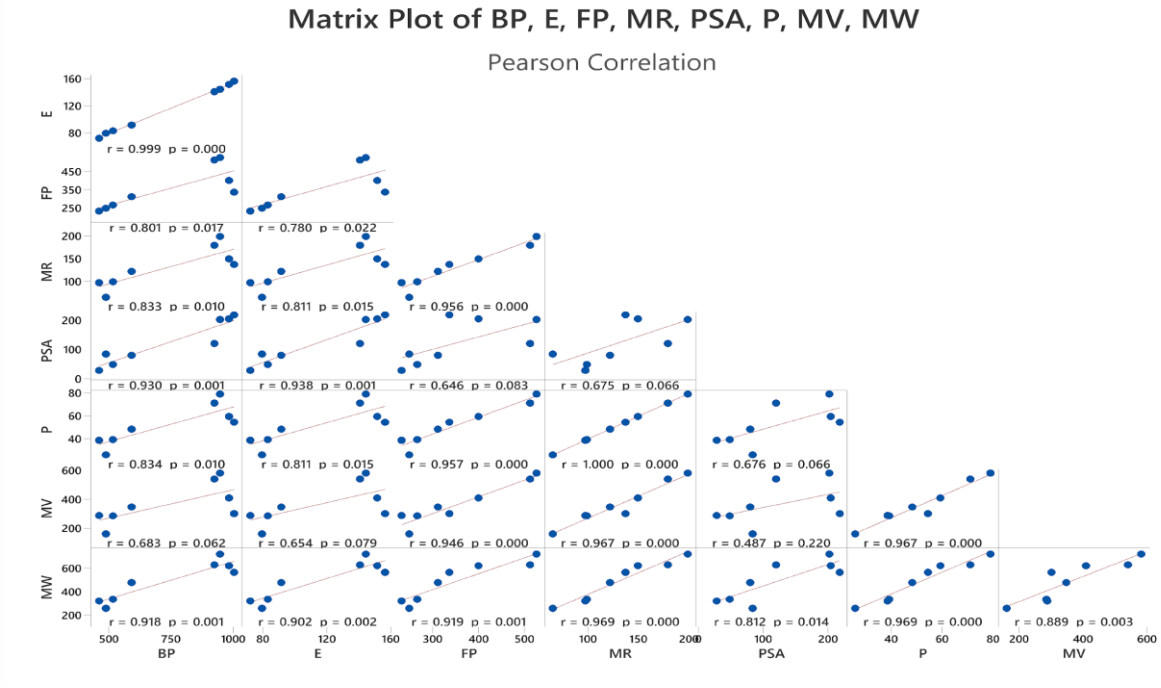


Figure 4: Pearson Correlation Matrix of Physicochemical Properties

Further, the linear relationships between the nine centrality measures ($DW(G)$, $DCW(G)$, $TCW(G)$, $CW(G)$, $LW(G)$, $HW(G)$, $HCW(G)$, $EW(G)$, $ECW(G)$) and the eight physicochemical properties (BP, E, FP, MR, PSA, P, MV, MW) were investigated. The full correlation matrix, detailing these relationships are presented in Table 7. This table served as a primary tool for identifying promising descriptors for each property. The individual contributions of each centrality measure to the prediction of specific physicochemical properties are further visualised through line charts of their respective Pearson correlation coefficients (Figure 5). Each subplot in Figure 5 illustrates the correlation strength of the nine centrality measures against a single physicochemical property, making it easy to graphically identify the best correlating descriptors for that specific property.

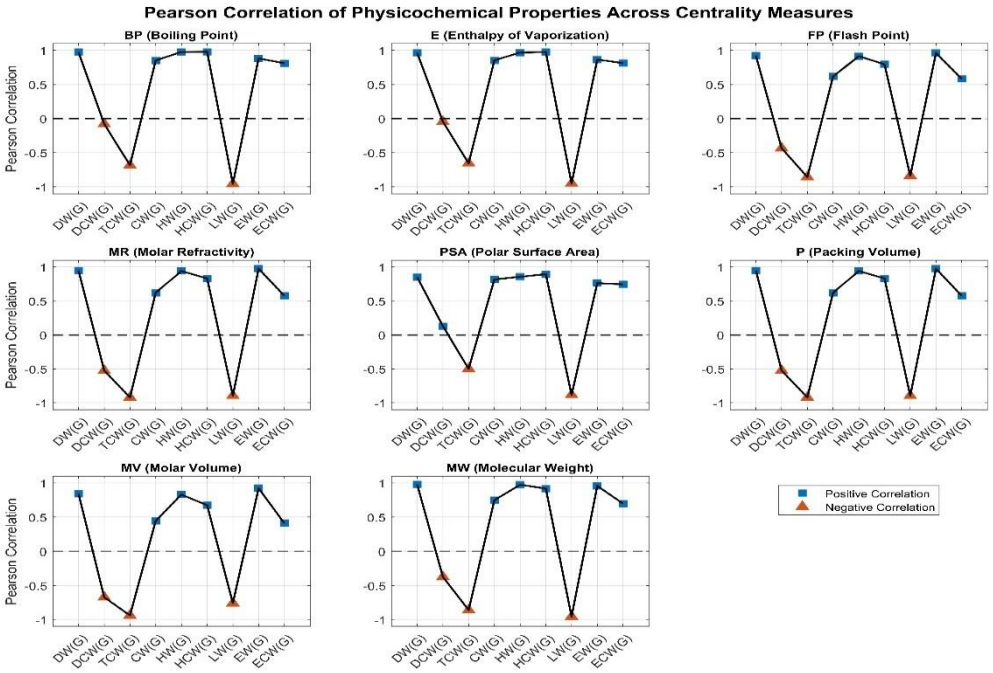


Figure 5: Pearson Correlation Coefficients of Molecular Descriptors with Individual Physicochemical Properties.

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XIV, Issue XII, December 2025

Table 7: Pearson Correlation Matrix of Centrality Measures and Physicochemical Properties

Measures	BP (mmHg)	E (kJ/mol)	FP (°C)	MR (cm^3)	PSA (Å²)	P (cm^3)	MV (cm^3)	MW (g/mol)
$DW(G)$	0.972	0.960	**0.921****	**0.949****	0.850	**0.949****	0.844	**0.978***
$DCW(G)$	-0.080	-0.046	-0.433	-0.525	0.129	-0.524	-0.673	-0.375
$TCW(G)$	-0.683	-0.655	-0.856	-0.920	-0.499	-0.920	**-0.939***	-0.860
$CW(G)$	0.851	0.853	0.615	0.621	0.817	0.621	0.445	0.749
$HW(G)$	**0.976****	**0.966****	0.914	0.941	0.856	0.941	0.831	**0.974****
$HCW(G)$	**0.978***	**0.974***	0.794	0.831	**0.894***	0.831	0.674	0.919
$LW(G)$	-0.956	-0.950	-0.839	-0.895	**-0.883****	-0.895	-0.760	-0.957
$EW(G)$	0.881	0.864	**0.959***	**0.975***	0.763	**0.975***	**0.923****	0.958
$ECW(G)$	0.810	0.812	0.582	0.577	0.746	0.577	0.411	0.698

Note: *Highest and **Second highest Pearson correlation values.

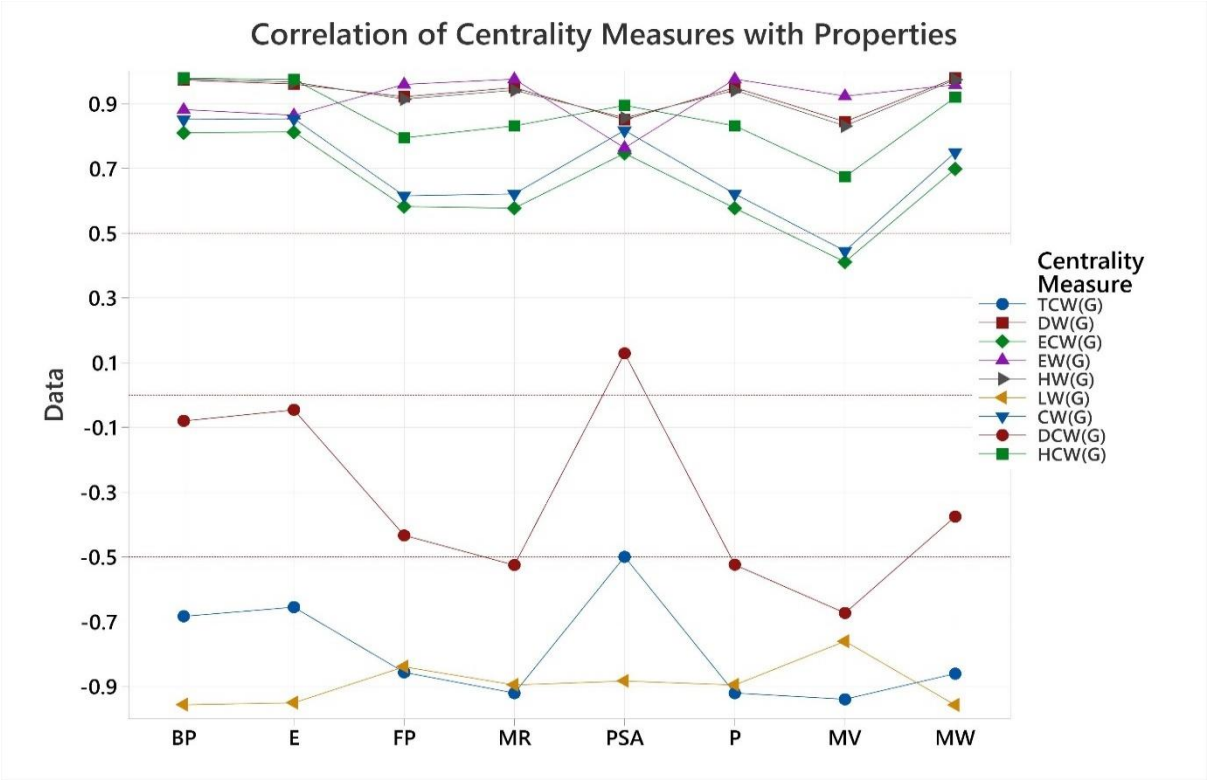


Figure 6: Overall Correlation Trends of Molecular Descriptors Across All Physicochemical Properties.

Beyond descriptor-property relationships, the inter-correlation among the centrality descriptors themselves was also rigorously assessed. Revealed instances of high collinearity. For instance, $DW(G)$ and $HW(G)$ exhibit a perfect positive correlation (r=1.000), indicating their redundancy within this dataset. The presence of such strong correlations among descriptors was a critical consideration in the subsequent model development phase, guiding the selection of single-predictor models to ensure robustness and avoid multicollinearity issues in the

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XIV, Issue XII, December 2025

final QSPR models. The overall trends and distribution of correlation coefficients across all properties for each centrality measure are summarised in Figure 6. This figure plots the range of correlation coefficients (from -1 to 1) for each of the nine centrality measures, with reference lines at ±0.5 and 0 to delineate strong positive, strong negative, and negligible correlations. As depicted, six measures ($DW(G)$, $CW(G)$, $HW(G)$, $HCW(G)$, $EW(G)$, $ECW(G)$) consistently show strong positive correlations across various properties, while $TCW(G)$ and $LW(G)$ primarily exhibit strong negative correlations. $DCW(G)$, however, generally displays very weak correlations, consistently falling close to the zero midline, indicating its limited predictive utility, as also observed numerically in Table 7. This comprehensive correlation analysis was crucial for identifying the most relevant descriptors that are central to developing robust linear QSPR models.

Regression Model Selection

Prior to model development, a comprehensive Pearson correlation analysis was conducted to assess the linear relationships between all physicochemical properties and molecular descriptors. This matrix served as an initial guide for identifying strong potential predictors. It also highlighted instances of high inter-correlation among certain descriptors, informing the subsequent decision to favour single-predictor models to ensure model robustness and avoid multicollinearity.

Following this comprehensive correlation analysis, which identified promising molecular descriptors (Table 7, Figure 5, and Figure 6), single-predictor linear regression models were developed for each of the eight physicochemical properties. The primary objective was to identify the most robust and statistically significant QSPR model for each property. To achieve this, all nine centrality measures were individually evaluated as potential predictors against each physicochemical property, resulting in 72 unique simple LRM's (linear regression models). Model performance was rigorously evaluated based on standard statistical parameters, including the coefficient of determination R^2, adjusted R^2 $\left(R^2_{adj}\right)$, the standard error of regression (S), F-statistic, and the model's p-value. Crucially, the predictive R^2 $\left(R^2_{pred}\right)$ was used as the primary criterion for assessing the model's ability to generalise to new, unseen data, reflecting its true predictive power.

Table 8: Summary of Optimised Single-Predictor QSPR Models for Physicochemical Properties.

Property	Optimal Descriptor	Regression Equation (Uncoded)	R-sq.	Adj. R-sq.	R-sq. (Pred)	S (Standard Error of Regression)	F-value	p-value (Model)
BP (mmHg)	$HCW(G)$	BP = -717 + 170.4 x $HCW(G)$	0.9526	0.9482	0.9358	52.4528	129.03	0.000
E (kJ/mol)	$HCW(G)$	E = -100.5 + 25.20 x $HCW(G)$	0.9489	0.9404	0.9256	8.3483	111.46	0.000
FP (ºC)	$EW(G)$	FP = 163.4 + 0.4512 x $EW(G)$	0.9205	0.9073	0.8613	35.5185	69.49	0.000
MR (cm^3)	$EW(G)$	MR = 55.92 + 0.1750 x $EW(G)$	0.9499	0.9416	0.9088	10.7617	113.86	0.000
PSA ($Å^2$)	$HCW(G)$	PSA = -308.0 + 51.0 x $HCW(G)$	0.7992	0.7657	0.6793	36.4949	23.87	0.003
P (cm^3)	$EW(G)$	P = 22.15 + 0.0694 x $EW(G)$	0.9507	0.9424	0.9104	4.2362	115.59	0.000

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XIV, Issue XII, December 2025

MV (cm^3)	$EW(G)$	MV = 139.8 + 0.5226 x $EW(G)$	0.8524	0.8278	0.7698	58.2660	34.65	0.001
MW (g/mol)	$DW(G)$	MW = 26.3 + 6.385 x $DW(G)$	0.9569	0.9497	0.9295	38.0859	133.05	0.000

Table 8 summarises the optimal single-predictor QSPR model identified for each physicochemical property. For each property, this table presents the descriptor that yielded the highest predictive performance R^2_{pred}, along with its corresponding regression equation in uncoded units and key statistical parameters. This consolidated view allows for a direct comparison of the best-performing models across all properties.

Validation of QSPR Models for Physicochemical Property Prediction

The coefficient of determination (R^2), Pearson's correlation coefficient, and root mean square error (RMSE) are critical metrics in the predicted versus, actual scatter plots (Figure 7), as they quantify the accuracy and reliability of the QSPR models for predicting physicochemical properties (BP, E, FP, MR, PSA, P, MV, MW) of eight drugs. R^2, derived from the square of Pearson corelation (r), indicates the proportion of variance in actual values explained by the model (e.g., R^2=0.837 for BP shows strong fit). r, calculated is the measure of linear correlation between actual and predicted values, with values near 1 indicating strong positive correlation. RMSE, computed as the square root of the mean squared differences between actual and predicted values, quantifies prediction error in the property's units, where lower values denote higher accuracy. These metrics, displayed on each plot, validate model performance, with tight clustering around the 45-degree line and high $\frac{R^2}{r}$, values for BP and E suggesting robust predictions, while larger RMSE values for PSA highlight areas for model refinement.

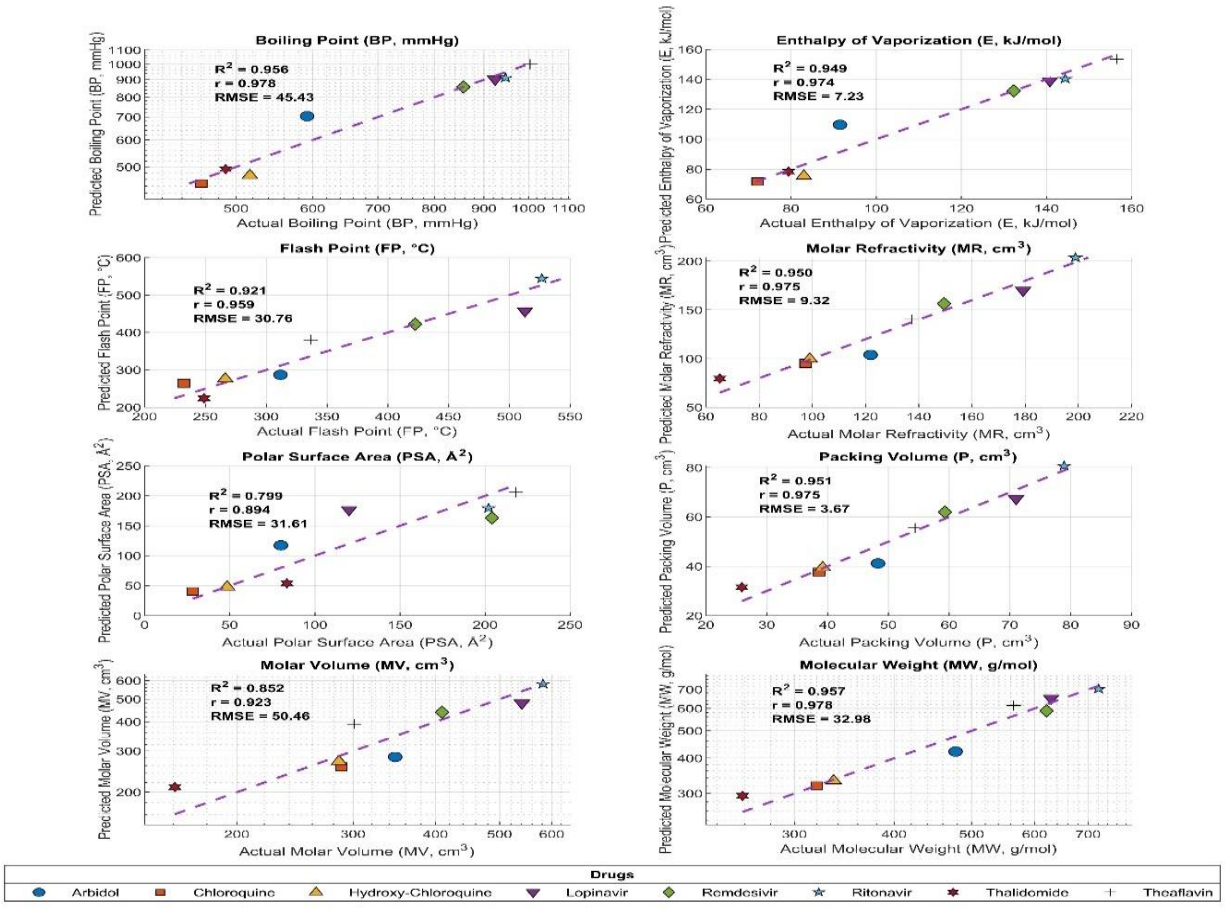


Figure 7. Scatter Plots of Predicted vs. Actual Physicochemical Properties for QSPR Model Validation, Using Marker Shapes with R^2, r and RMSE.

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XIV, Issue XII, December 2025

RESULTS

Imputation Outcomes

Missing experimental values for Boiling Point (BP), Enthalpy of Vaporisation (E), and Flash Point (FP) for remdesivir were successfully imputed using correlation-based simple linear regression models. The selection of descriptors for this imputation was guided by strong Pearson correlation coefficients: $HCW(G)$ exhibited a correlation of 0.976 with BP and 0.973 with E, while $EW(G)$ showed a correlation of 0.957 with FP, as detailed in Table 4.

Using remdesivir's known descriptor values, the following linear regression equations were applied:

- For Boiling Point (BP): BP = -717 + 170.4 × $HCW(G)$
- For Enthalpy of Vaporisation (E): E = -100.5 + 25.20 × $HCW(G)$
- For Flash Point (FP): FP = 163.4 + 0.4512 × $EW(G)$

Applying remdesivir's $HCW(G)$ value of 9.238375, its estimated Boiling Point was 857.22 mmHg and its estimated Enthalpy of Vaporisation was 132.31 kJ/mol. For Flash Point, using an EW value of 574, the estimated value was 422.39 °C. These imputed values, essential for completing the dataset for subsequent QSPR modelling, are summarised in Table 5. Visual representations of the regression models used for these imputations are presented in Figure 3.

Data Characteristics and Correlations

The fundamental descriptive statistics for all 17 centrality measures and 8 physicochemical properties across the eight drug molecules, including remdesivir, are presented in Table 6. This summary illustrates the inherent variability and ranges within the dataset, highlighting significant scale disparities.

A comprehensive Pearson correlation analysis was conducted on the complete dataset (post missing data substitution) to elucidate the linear relationships between all variables. Figure 4 provides a Pearson correlation matrix of the physicochemical properties themselves, showing strong intercorrelations such as that between Boiling Point (BP) and Enthalpy of Vaporisation (E).

The full correlation matrix detailing the relationships between the nine centrality measures and the eight physicochemical properties is presented in Table 7. This matrix was instrumental in identifying promising descriptors for QSPR modelling. Key findings include:

- Strong Positive Correlations: $HCW(G)$ exhibited strong positive correlations with BP (0.978), E (0.974), and PSA (0.894). EW demonstrated high positive correlations with FP (0.959), MR (0.975), P (0.975), and MV (0.923). $DW(G)$ showed a substantial positive correlation with MW (0.978).
- Strong Negative Correlations: Conversely, $TCW(G)$ and $LW(G)$ consistently displayed pronounced negative correlations across multiple properties. Notably, $TCW(G)$ was strongly associated with MV (-0.939), MR (-0.920), and P (-0.920), while $LW(G)$ correlated negatively with BP (-0.956), E (-0.950), and MW (-0.957).
- Weak Correlations: $DCW(G)$ generally displayed very weak correlations, consistently close to zero, indicating limited predictive utility for the studied properties.

The individual contributions and correlation strengths of each centrality measure against specific physicochemical properties are visually represented through line charts of their respective Pearson correlation coefficients in Figure 5. Figure 6 further summarises the overall trends and distribution of correlation coefficients

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XIV, Issue XII, December 2025

for each centrality measure across all properties, highlighting those consistently showing strong positive or negative associations.

Furthermore, a rigorous assessment of the intercorrelation among the centrality measure descriptors was conducted. This analysis identified instances of pronounced collinearity, exemplified by a perfect positive correlation (r=1.000) between $DW(G)$ and $HW(G)$, patterns of which are visually represented in the comprehensive correlation matrix of Figure 4. The presence of such strong inter-descriptor correlations significantly guided the selection of single-predictor models for QSPR development, a strategy adopted to ensure model stability and effectively avoid issues associated with multicollinearity.

Optimised QSPR Model Performance

The overall predictive capabilities of the single-predictor QSPR models, assessed by the predictive R^2 $\left(R^2_{pred}\right)$, are comprehensively presented in Table 9. This table highlights the R^2_{pred} value for each of the 72 individual models (9 centrality measures × 8 physicochemical properties), providing a robust measure of their ability to generalise to unseen data through LOO (leave-one-out) cross-validation. Values closer to 1 indicate higher predictive accuracy. Notably, several descriptors consistently yield high R^2_{pred} values across various properties, affirming their robust predictive potential. For instance, the $HCW(G)$ measure shows exceptional predictive power for BP (R^2_{pred}=0.9358) and E (R^2_{pred}=0.9256), while EW(G) demonstrates strong predictive performance for FP (R^2_{pred}=0.8613), MR (R^2_{pred}=0.9088), and PV (R^2_{pred}=0.9104). $DW(G)$ also exhibits high predictive accuracy for MW (R^2_{pred}=0.9295).

Table 9: Predictive R^2 $\left(R^2_{pred}\right)$ Values for Single-Predictor QSPR Models of Physicochemical Properties.

Drugs	$DW(G)$	$DCW(G)$	$TCW(G)$	$CW(G)$	$HW(G)$	$HCW(G)$	$LW(G)$	$EW(G)$	$ECW(G)$
BP (mmHg)	0.9008	0.0000	0.2310	0.5943	**0.9164****	**0.9358***	0.8594	0.5803	0.5037
E (kJ/mol)	0.8633	0.0000	0.1817	0.5942	**0.881****	**0.9256***	0.8398	0.5353	0.4974
FP (ºC)	**0.7428****	0.0000	0.2832	0.0000	0.7145	0.5689	0.4921	**0.8613***	0.0000
MR (cm^3)	**0.8158****	0.0000	0.7305	0.2210	0.7879	0.3945	0.6103	**0.9088***	0.0017
PSA ($Å^2$)	0.5146	0.0000	0.0000	0.5249	0.5322	**0.6793***	**0.6292****	0.3403	0.3540
P (cm^3)	**0.8163****	0.0000	0.7287	0.2190	0.7883	0.3945	0.6103	**0.9104***	0.0016
MV (cm^3)	0.4899	0.0833	**0.7433****	0.0000	0.4438	0.0000	0.1926	**0.7698***	0.0000
MW (g/mol)	**0.9295***	0.0000	0.6031	0.3081	**0.9143****	0.6897	0.8427	0.8455	0.2540

Note: *Highest and **Second highest predictive efficiency

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XIV, Issue XII, December 2025

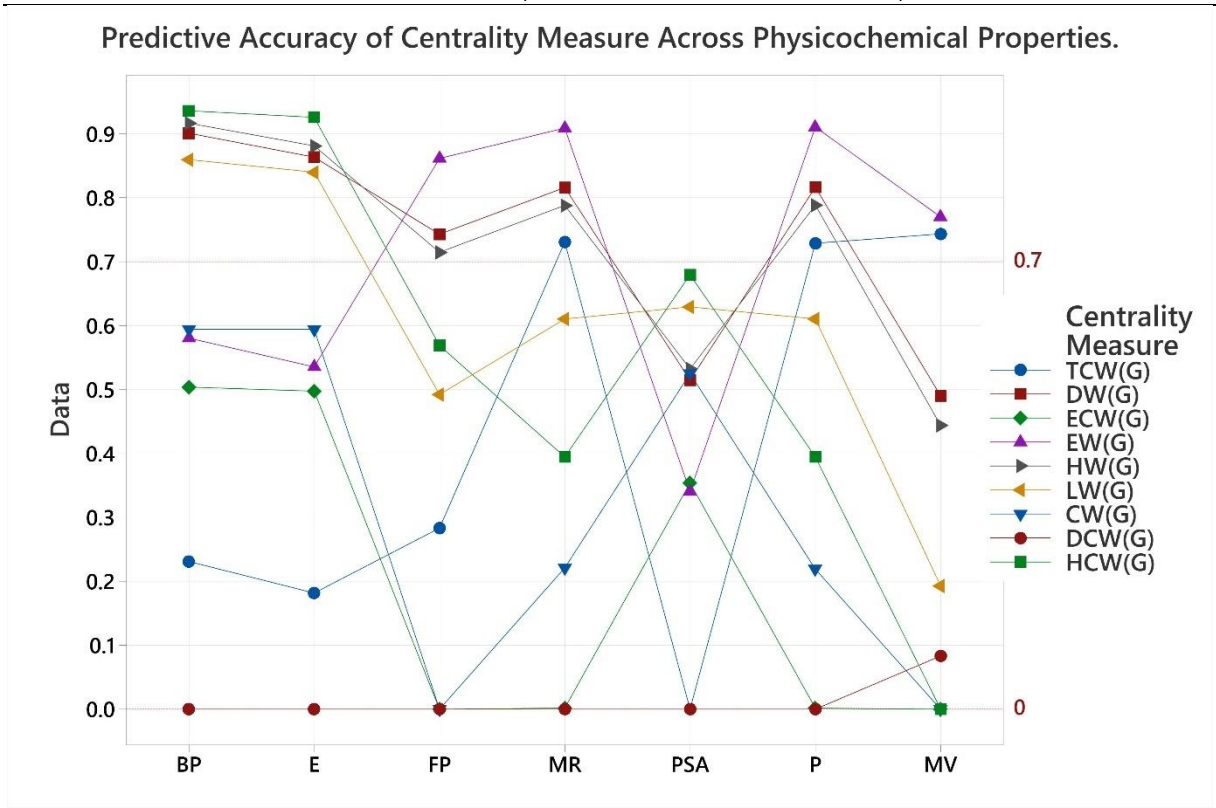


Figure 8. Comparative Predictive Performance (R^2_{pred}) of Individual Centrality Measures for Each Physicochemical Property.

A visual representation of these predictive performances is provided in Figure 8, which comparatively plots the R^2_{pred} values for all centrality measures across each physicochemical property. This figure clearly delineates the best-performing descriptors for each property and illustrates the relative predictive strengths and weaknesses of each centrality measure. Notably, the consistently low R^2_{pred} values observed for $DCW(G)$ across most properties, as well as for $TCW(G)$ and $LW(G)$ in certain cases, further underscore their limited predictive utility in single-predictor models for this dataset. Conversely, measures like $HCW(G)$ and $EW(G)$ frequently occupy the higher end of the predictive spectrum, reinforcing their value as robust QSPR descriptors. The comparison between the model's fit (R^2) and its predictive ability (R^2_{pred}) is critical, as a high R^2_{pred} confirms that the observed fit is not merely due to overfitting but reflects true model robustness & generalisability.

DISCUSSION

Effectiveness of Missing Data Imputation

The imputation of missing experimental data for remdesivir's Boiling Point (BP), Enthalpy of Vaporisation (E), and Flash Point (FP) was critical for comprehensive QSPR model development. By employing correlation-based linear regression with highly correlated centrality measures, accurate estimated values were generated. This efficient approach ensured dataset completeness without data exclusion, a crucial advantage for a limited sample size. The success of this missing data substitution method enhances the robustness and applicability of the subsequent QSPR analysis, demonstrating a valuable strategy for handling incomplete datasets in cheminformatics studies, which is consistently evident from robust high R^2_{pred}.

Interpretation of Developed QSPR Models

The comprehensive correlation analysis illuminated strong linear relationships between molecular graphs, represented by the nine centrality measures, and the eight physicochemical properties. As presented in Table 7 and Figure 5, HCW(G) and $EW(G)$ consistently showed strong positive correlations ($r \geq 0.9$) with properties like Boiling Point, Enthalpy of Vaporisation, Flash Point, Molar Refractivity, Packing Volume, and Molar

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XIV, Issue XII, December 2025

Volume. This suggests these measures effectively capture structural features influencing intermolecular forces and bulk properties. Conversely, $TCW(G)$ and $LW(G)$ frequently exhibited strong inverse relationships ($r \leq -0.8$) with properties such as Molar Volume and Molecular Weight, indicating their sensitivity to structural attributes inversely related to these properties.

The selection of optimal single-predictor linear regression models, based on their predictive R^2 (R^2_{pred}) derived from LOO cross-validation (Table 8, Table 9), confirmed the robust predictive power and generalisability of these relationships. The high R^2_{pred} Values demonstrate that even a single, well-chosen centrality measure can establish highly accurate and interpretable QSPR models for this dataset, elucidating the structural determinants of physicochemical behaviour.

Limitations and Future Directions

A primary limitation of this study is the small sample size of eight drug compounds. While rigorous cross-validation using PRESS and the focus on predictive R^2 mitigate overfitting, the small observations DOF (Degree of Freedom, N-1=7) restricts the complexity of models that can be reliably developed and limits the capture of more intricate, multi-descriptor relationships (MLR Models). The high collinearity observed among certain descriptors further dictated the focus on single-predictor models for robustness because of the small sample size in model building.

Future research should aim to expand the dataset with a larger and more structurally diverse set of compounds. This would facilitate the exploration of advanced QSPR methodologies, including:

- Multiple Linear Regression (MLR): Incorporating multiple, non-redundant centrality measures to potentially improve accuracy and mechanistic interpretability.
- Non-linear and Machine Learning Models: Investigating algorithms capable of capturing complex, non-linear structure-property relationships.
- Expanded Descriptor Sets: Exploring additional classes of molecular descriptors to provide complementary structural information.

These advancements would refine predictive capabilities and deepen the understanding of molecular mechanisms, thereby facilitating more efficient rational drug design.

CONCLUSIONS

This study has successfully developed and evaluated Quantitative Structure-Property Relationship (QSPR) models for eight physicochemical properties across a diverse set of eight drug compounds related to COVID-19 treatment. A critical aspect of this research involved efficiently addressing missing data; the imputation of Boiling Point, Enthalpy of Vaporisation, and Flash Point for remdesivir drug using correlation-based linear regression proved to be an optimally accurate and effective, ensuring a complete and robust dataset for subsequent statistical analyses.

The comprehensive correlation analysis performed underscores the significant influence of molecular graphs on diverse physicochemical properties, which are the basis for QSPR modelling. Statistical analysis demonstrates that specific molecular graph-based centrality measures, particularly $HCW(G)$, $EW(G)$ serve as powerful descriptors, exhibiting strong positive correlations with key properties such as Boiling Point, Enthalpy of Vaporisation, Molar Refractivity, and Packing Volume. Conversely, measures like $TCW(G)$ and $LW(G)$ consistently show strong inverse relationships with several properties, including Molar Volume and Molecular Weight. These identified relationships are pivotal for selecting optimal descriptors in QSPR model development.

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XIV, Issue XII, December 2025

The developed single-predictor linear regression models, summarised in Table 8 provide interpretable relationships between molecular structure and physicochemical behaviour. The rigorous validation employing the PRESS (Predicted Residual Error Sum of Squares) statistic derived from LOOCV (leave-one-out cross-validation) ensures the high predictive accuracy and generalisability of these models for external data. These models not only advance our fundamental understanding of structure-property relationships but also offer detailed methodology & use of practical tools for predicting properties of relevant new compounds. By elucidating the structural determinants of physicochemical behaviour, this research contributes valuable insights for rational drug design and molecular engineering, facilitating more efficient and targeted development processes in cheminformatics.

FUNDING

The authors declare that no funds, grants, or other support were received during the preparation of this manuscript.

ACKNOWLEDGMENTS

The author, Pavithra M. sincerely acknowledges the financial support provided by the Department of Science and Technology (DST) – Karnataka Science and Technology Promotion Society (KSTePS), Government of Karnataka (Ref. No. MP02/2023-24/430, dated 23rd January 2024), through the DST-KSTePS Fellowship.

Disclosure statement

The authors report there are no competing interests to declare.

Author Contributions

Pavithra M.: Conceptualization, Methodology, Data Curation, Formal Analysis, Literature Review, Original Draft preparation and Editing; Veena Mathad: Supervision, writing-review and editing.

REFERENCES

1. Zhu, N., D. Zhang, W. Wang, X. Li, B. Yang, J. Song, X. Zhao, B. Huang, W. Shi and R. Lu, A novel coronavirus from patients with pneumonia in China, 2019. New England journal of medicine. **382** (2020), no. 8, 727-733, DOI 10.1056/NEJMoa2001017
2. Organization, W. H. WHO Director-General's opening remarks at the media briefing on COVID-19. 2020.
3. Sanders, J. M., M. L. Monogue, T. Z. Jodlowski and J. B. Cutrell, Pharmacologic treatments for coronavirus disease 2019 (COVID-19): a review. Jama. **323** (2020), no. 18, 1824-1836, DOI 10.1001/jama.2020.6019
4. Wang, Y., D. Zhang, G. Du, R. Du, J. Zhao, Y. Jin, S. Fu, L. Gao, Z. Cheng and Q. Lu, Remdesivir in adults with severe COVID-19: a randomised, double-blind, placebo-controlled, multicentre trial. The lancet. **395** (2020), no. 10236, 1569-1578, DOI 10.1016/S0140-6736(20)31022-9
5. Yao, X., F. Ye, M. Zhang, C. Cui, B. Huang, P. Niu, X. Liu, L. Zhao, E. Dong and C. Song, In vitro antiviral activity and projection of optimized dosing design of hydroxychloroquine for the treatment of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). Clinical infectious diseases. **71** (2020), no. 15, 732-739, DOI 10.1093/cid/ciaa237
6. Blaising, J., S. J. Polyak and E.-I. Pécheur, Arbidol as a broad-spectrum antiviral: an update. Antiviral research. **107** (2014), no. 84-94, DOI 10.1016/j.antiviral.2014.04.006
7. Savarino, A., J. R. Boelaert, A. Cassone, G. Majori and R. Cauda, Effects of chloroquine on viral infections: an old drug against today's diseases. The Lancet infectious diseases. **3** (2003), no. 11, 722-727, DOI 10.1016/S1473-3099(03)00806-5
8. Gagliardini, R., A. Cozzi-Lepri, A. Mariano, F. Taglietti, A. Vergori, A. Abdeddaim, F. Di Gennaro, V. Mazzotta, A. Amendola and G. D'Offizi, No efficacy of the combination of lopinavir/ritonavir plus

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XIV, Issue XII, December 2025

hydroxychloroquine versus standard of care in patients hospitalized with COVID-19: a non-randomized comparison. Frontiers in Pharmacology. **12** (2021), no. 621676, DOI 10.3389/fphar.2021.621676
9. Cao, B., Y. Wang, D. Wen, W. Liu, J. Wang, G. Fan, L. Ruan, B. Song, Y. Cai and M. Wei, A trial of lopinavir–ritonavir in adults hospitalized with severe Covid-19. New England journal of medicine. **382** (2020), no. 19, 1787-1799, DOI 10.1056/NEJMoa2001282
10. Franks, M. E., G. R. Macpherson and W. D. Figg, Thalidomide. The Lancet. **363** (2004), no. 9423, 1802-1811, DOI 10.1016/S0140-6736(04)16308-3
11. Zu, M., F. Yang, W. Zhou, A. Liu, G. Du and L. Zheng, In vitro anti-influenza virus and anti-inflammatory activities of theaflavin derivatives. Antiviral research. **94** (2012), no. 3, 217-224, DOI 10.1016/j.antiviral.2012.04.001
12. Beigel, J. H., K. M. Tomashek, L. E. Dodd, A. K. Mehta, B. S. Zingman, A. C. Kalil, E. Hohmann, H. Y. Chu, A. Luetkemeyer and S. Kline, Remdesivir for the treatment of Covid-19—preliminary report. New England Journal of Medicine. **383** (2020), no. 19, 1813-1836, DOI 10.1056/NEJMoa2007764
13. De Clercq, E., Anti-HIV drugs: 25 compounds approved within 25 years after the discovery of HIV. International journal of antimicrobial agents. **33** (2009), no. 4, 307-320, DOI 10.1016/j.ijantimicag.2008.10.010
14. Sheahan, T. P., A. C. Sims, R. L. Graham, V. D. Menachery, L. E. Gralinski, J. B. Case, S. R. Leist, K. Pyrc, J. Y. Feng and I. Trantcheva, Broad-spectrum antiviral GS-5734 inhibits both epidemic and zoonotic coronaviruses. Science translational medicine. **9** (2017), no. 396, eaal3653, DOI 10.1126/scitranslmed.aal3653
15. Hung, I. F.-N., K.-C. Lung, E. Y.-K. Tso, R. Liu, T. W.-H. Chung, M.-Y. Chu, Y.-Y. Ng, J. Lo, J. Chan and A. R. Tam, Triple combination of interferon beta-1b, lopinavir–ritonavir, and ribavirin in the treatment of patients admitted to hospital with COVID-19: an open-label, randomised, phase 2 trial. The lancet. **395** (2020), no. 10238, 1695-1704, DOI 10.1016/S0140-6736(20)31042-4
16. Todeschini, R. and V. Consonni, Molecular descriptors for chemoinformatics: volume I: alphabetical listing/volume II: appendices, references; John Wiley & Sons, 2009.
17. Ivanciuc, O., S. L. Taraviras and D. Cabrol-Bass, Quasi-orthogonal basis sets of molecular graph descriptors as a chemical diversity measure. Journal of Chemical Information and Computer Sciences. **40** (2000), no. 1, 126-134, DOI 10.1021/ci990064x
18. Todeschini, R. and V. Consonni, Handbook of molecular descriptors; John Wiley & Sons, 2008. DOI: 10.1002/9783527613106.
19. Ivanciuc, O., Chemical graphs, molecular matrices and topological indices in chemoinformatics and quantitative structure-activity relationships §. Current computer-aided drug design. **9** (2013), no. 2, 153-163, DOI 10.2174/1573409911309020002
20. Kara, Y., Y. S. Özkan, A. Ullah, Y. S. Hamed and M. B. Belay, QSPR modeling of some COVID-19 drugs using neighborhood eccentricity-based topological indices: A comparative analysis. PLoS One. **20** (2025), no. 5, e0321359, DOI 10.1371/journal.pone.0321359
21. Jyothish, K. and R. Santiago, Quantitative Structure–Property Relationship Modeling with the Prediction of Physicochemical Properties of Some Novel Duchenne Muscular Dystrophy Drugs. ACS omega. **10** (2025), no. 4, 3640, DOI 10.1021/acsomega.4c08572
22. Tiikkainen, P., L. Bellis, Y. Light and L. Franke, Estimating error rates in bioactivity databases. Journal of chemical information and modeling. **53** (2013), no. 10, 2499-2505, DOI 10.1021/ci400099q
23. Little, R. J. and D. B. Rubin, Statistical analysis with missing data; John Wiley & Sons, 2019. DOI: 10.1002/9781119482260.
24. Schneider, T., Analysis of incomplete climate data: Estimation of mean values and covariance matrices and imputation of missing values. Journal of climate. **14** (2001), no. 5, 853-871, DOI 10.1175/1520-0442(2001)014%3C0853:AOICDE%3E2.0.CO;2
25. Li, J., S. Guo, R. Ma, J. He, X. Zhang, D. Rui, Y. Ding, Y. Li, L. Jian and J. Cheng, Comparison of the effects of imputation methods for missing data in predictive modelling of cohort study datasets. BMC Medical Research Methodology. **24** (2024), no. 1, 41, DOI 10.1186/s12874-024-02173-x
26. Alwateer, M., E.-S. Atlam, M. M. Abd El-Raouf, O. A. Ghoneim and I. Gad, Missing data imputation: A comprehensive review. Journal of Computer and Communications. **12** (2024), no. 11, 53-75, DOI 10.4236/jcc.2024.1211004

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XIV, Issue XII, December 2025

27. Schneiderman, E. D., C. J. Kowalski and S. M. Willis, Regression imputation of missing values in longitudinal data sets. International journal of bio-medical computing. **32** (1993), no. 2, 121-133, DOI 10.1016/0020-7101(93)90051-7
28. Sun, Y., J. Li, Y. Xu, T. Zhang and X. Wang, Deep learning versus conventional methods for missing data imputation: A review and comparative study. Expert Systems with Applications. **227** (2023), no. 120201, DOI 10.1016/j.eswa.2023.120201
29. Allen, D. M., The relationship between variable selection and data agumentation and a method for prediction. technometrics. **16** (1974), no. 1, 125-127, DOI 10.1080/00401706.1974.10489157
30. Myers, R. H., Classical and modern regression with applications; Boston : PWS-Kent, 1990.
31. Golbraikh, A. and A. Tropsha, Beware of q2! Journal of molecular graphics and modelling. **20** (2002), no. 4, 269-276, DOI 10.1016/S1093-3263(01)00123-1
32. Roy, K., S. Kar and R. N. Das, Understanding the basics of QSAR for applications in pharmaceutical sciences and risk assessment; Academic press, 2015.
33. Consonni, V., D. Ballabio and R. Todeschini, Evaluation of model predictive ability by external validation techniques. Journal of chemometrics. **24** (2010), no. 3-4, 194-201, DOI 10.1002/cem.1290
34. Pavel, H., A. Roy, A. Santra and S. Chakravarthy. Degree centrality definition, and its computation for homogeneous multilayer networks using heuristics-based algorithms. In International Joint Conference on Knowledge Discovery, Knowledge Engineering, and Knowledge Management, 2022; Springer: pp 28-52.
35. Mathad, V. and M. Pavithra, Closeness Centrality Weight and Edge Closeness Centrality Weight of Graphs. Dynamics of Continuous, Discrete and Impulsive Systems Series B: Applications & Algorithms. **32** (2025), no. 109-124, DOI
36. Sukumaran, S. and S. Unnithan, Mathematical Perspectives of Leverage Centrality: A Review. Indian Journal of Science and Technology. **16** (2023), no. 39, 3325-3331, DOI 10.17485/IJST/v16i39.1234
37. Berberler, M. E., Leverage centrality analysis of infrastructure networks. Numerical Methods for Partial Differential Equations. **37** (2021), no. 1, 767-781, DOI 10.1002/num.22551
38. Ortega, J. M. E. and R. G. Eballe, Harmonic Centrality and Centralization of Some Graph Products. arXiv preprint arXiv:2205.03791. (2022), no. DOI 10.9734/ARJOM/2022/v18i530377
39. Hage, P. and F. Harary, Eccentricity and centrality in networks. Social networks. **17** (1995), no. 1, 57-63, DOI 10.1016/0378-8733(94)00248-9