Application of the "EigenValue Analysis (EVANS)" Methodology to Build Quantitative Structure Pharmacokinetic Relationship Models

Anish Gomatam[¥], Blessy Joseph[¥], Poonam Advani[₸], Mushtaque Shaikh[ǂ], Evans Coutinho[¥*]

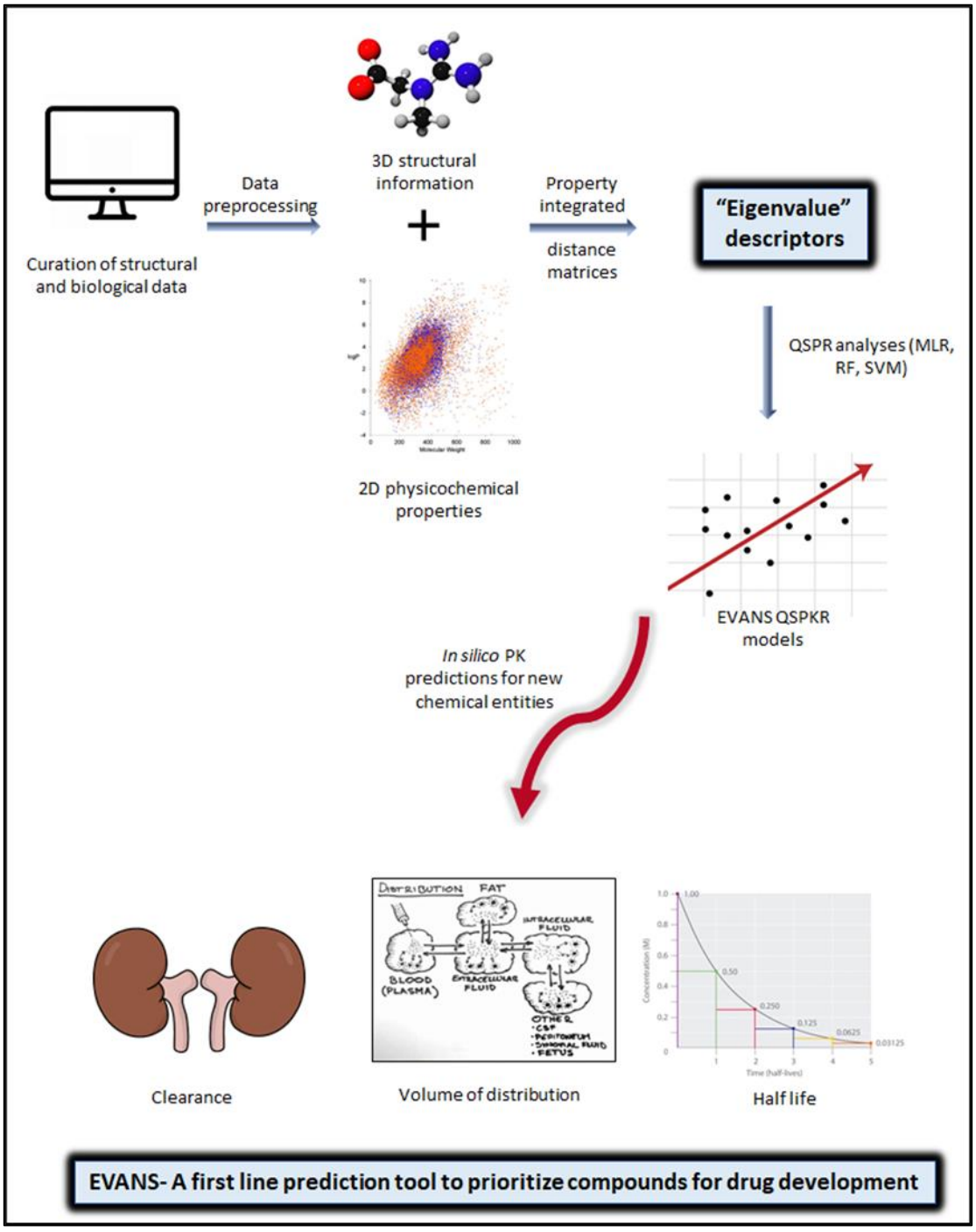
Molecular Simulations Group, Department of Pharmaceutical Chemistry, Bombay College of Pharmacy, Mumbai, India

[₸]*Mumbai Educational Trust's Institute of Pharmacy, Mumbai, India*

[ǂ]*Department of Pharmaceutical Chemistry, Vivekanand Education Society's College of Pharmacy, Mumbai, India*

*Correspondence E-mail: evans.coutinho@bcp.edu.in

Graphical Abstract



ABSTRACT

Quantitative structure-property relationship (QSPR) modelling has been a cornerstone of *in silico* drug development for several decades. The traditional Hansch methods have evolved into Quantitative structure-activity relationships (QSARs) in multiple dimensions, however, drawbacks remain. Many 3D-QSAR methods are heavily reliant on accurate alignment of low energy molecular structures and cannot treat drugs with multiple chiral centres which are often biologically screened as racemates. We present EigenValue ANalySis (EVANS), a QSPR methodology that considers 3D molecular information of enantiomeric ensembles of chiral molecules without the need to perform an alignment step. EVANS follows an intricate molecular modelling protocol that generates orthogonal eigenvalues from hybrid matrices of physicochemical properties and 3D structure; these eigenvalues are used as independent variables in QSPR analyses. The EVANS formalism has been presented and deployed to build quantitative structure pharmacokinetic relationship (QSPKR) models on a benchmark dataset for three critical PK parameters: steady-state volume of distribution (VD_{ss}), clearance (CL), and half-life ($t_{1/2}$). Predictive QSPKR models were built by using the eigenvalues generated via the EVANS methodology in conjunction with multiple linear regression (MLR), random forest (RF), and support vector machine (SVM) algorithms, and it was observed that the EVANS QSPKR models sync with published work in the literature. Thus, we present the EVANS methodology as a first-line prediction tool to prioritise compounds in drug discovery and development.

Keywords: Eigenvalues, QSAR, QSPR, pharmacokinetics, computational ADME, chemometrics, machine learning

INTRODUCTION

Pioneered by Corwin Hansch in the 1960s, Quantitative Structure-Activity Relationships (QSARs) and Quantitative Structure-Property Relationships (QSPRs) have been extensively studied for several years. Hansch pioneered the work in this area with his studies of the herbicidal effects of phenylacetic acids using multiple linear regression (MLR), an approach that is relevant even today owing to its simplicity and ease of interpretability[1]. Since then, QSARs have evolved tremendously, both in terms of the dimensionality and chemometric methods, with the traditional Hansch approach being supplemented with "black-box" machine learning and deep learning methods coupled with increasing dimensionality: 3D, 4D, 5D, and even 6D-QSAR. The advantages of multidimensional QSAR are obvious; the consideration of 3D structure and stereochemistry, which its 2D counterpart simply cannot do. However, this comes with its pitfalls, 3D-QSARs are heavily reliant on accurate alignment of low energy structures for descriptor calculation. Moreover, 3D-QSAR makes a very important assumption: the lowest energy conformation of the molecule is the bioactive conformation, and it is this single conformation that is responsible for the biological effect[2]. This does not hold true, especially in the case of molecules with multiple chiral centres where the physiological response may be due to additive contributions of multiple enantiomeric states. This consideration of 3D structure and the existence of multiple enantiomeric states assumes greater importance when considering the pharmacokinetics (PK) of drugs, as the various processes in drug disposition involve interactions between chiral drugs and chiral biological macromolecules[3].

PK, in medicinal chemistry parlance, is loosely defined as 'what the body does to the drug' and refers to the time course of the drug's absorption, distribution, metabolism, and excretion from the body[4]. It is evident that in addition to efficacy, any molecule that is intended for therapeutic use must possess desirable PK attributes to accumulate in sufficient concentration at the site of action while not causing harmful side effects. The two primary PK parameters which have major clinical significance are the steady-state volume of distribution and (VD_{ss}) and clearance (CL)[5]. VD_{ss} is a proportionality constant that represents a molecule's propensity to remain in the plasma or redistribute into tissue compartments. CL describes how efficiently a drug is removed from the body and determines the maintenance dose of the drug, much like $VD_{ss,}$ is used as a determinant of the loading dose[6]. These primary PK parameters can be used for the calculation of $t_{1/2}$, which is a measure of the time a drug stays in the body[5].

This work aims to model the PK properties of drugs by defining a QSPR methodology titled 'EigenValue ANalySis', abbreviated as EVANS. The uniqueness of the EVANS methodology is that it incorporates 3D information and accounts for the contribution of multiple chiral states towards a particular biological endpoint without the need to perform an alignment procedure. This methodology has previously been benchmarked on pharmacodynamic datasets with encouraging results[7]. We now deploy the EVANS methodology to build predictive Quantitative Structure Pharmacokinetic

Relationship (QSPKR) models using human intravenous clinical PK data. We discuss the EVANS methodology in detail and report EVANS models enumerating three PK parameters: VD_{ss}, CL, and $t_{1/2}$.

EXPERIMENTAL SECTION

Datasets

One of the largest datasets in the public domain with human clinical data was selected for modelling PK properties. The dataset was pruned by removing molecules with missing or ambiguous values for the PK parameters, compounds having more than 7 chiral centres, high molecular weight entities (proteins and peptides), monoclonal antibodies, lanthanides, and drugs containing transition state metals; this resulted in a final dataset of 474 molecules. The dataset is structurally and biologically diverse and contains molecules of various chemical classes. The EVANS formalism was run on this dataset and QSPKR models were built for three PK parameters: VD_{ss}, CL, and $t_{1/2}$. The distribution of these properties is given in **Table 1**.

Table 1: Distribution of PK parameters used in this work

Parameter	**log VD_{ss}**	**log CL**	**log $t_{1/2}$**
Range	1.6 to 5.84	-2.43 to 2.46	-0.92 to 3.1
Mean	3.02	0.56	0.67

Methodology Application- A Blueprint

An important element in EVANS formalism is the concept of hybrid descriptors that encode both 3D structural and physicochemical information for modelling QSPR properties. The steps in the EVANS methodology are shown in **Figure 1**. The formalism has been published earlier[7], however, some aspects have been refined for use in QSPR modelling. We succinctly explain the workings of the EVANS formalism below.

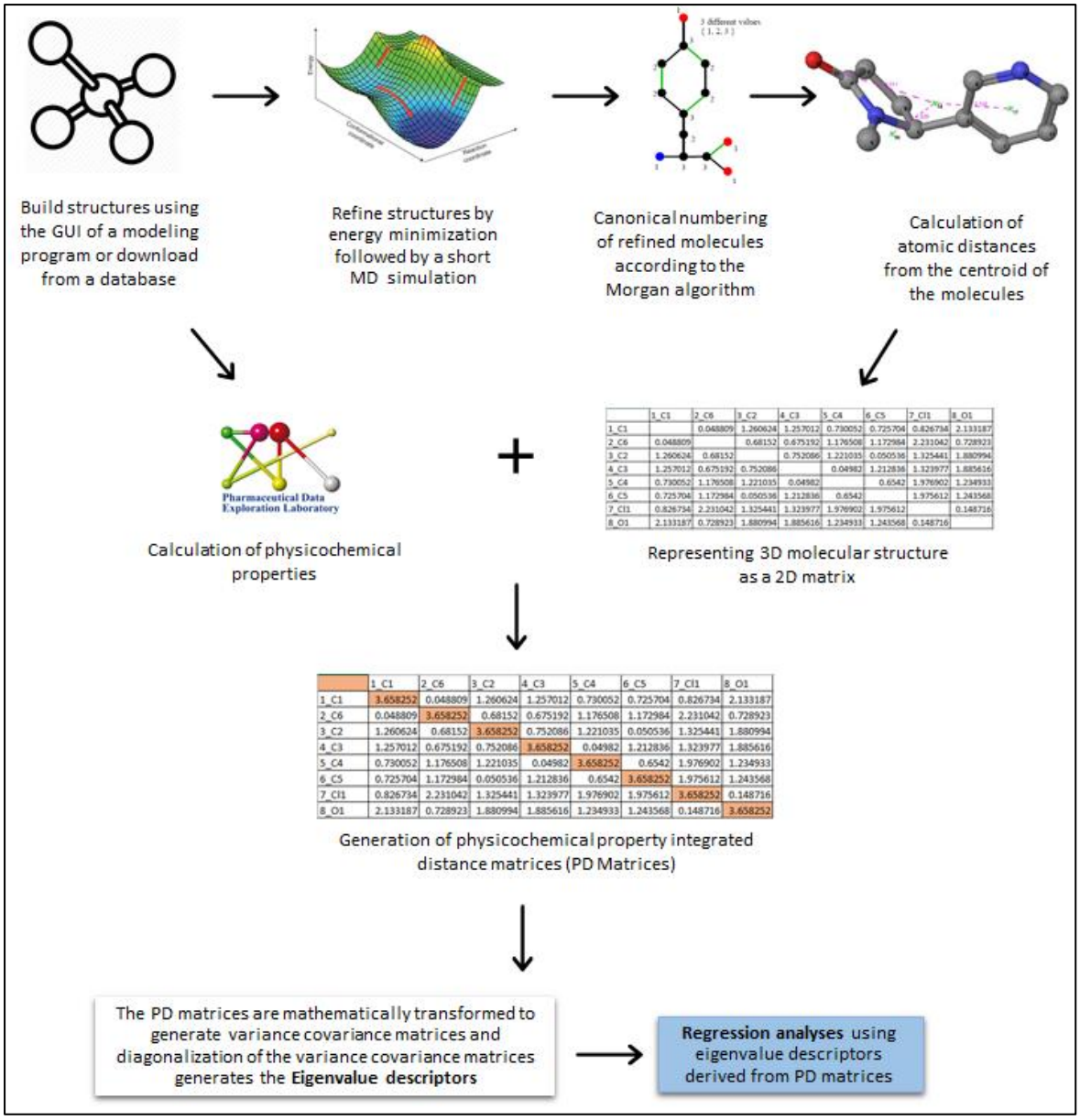


Figure 1: A blueprint of the EVANS methodology

Broadly, the methodology can be divided into three parts.

1. ***Representation of molecular structure***

   Three-dimensional structures of molecules in the dataset are projected onto a square 2D matrix. This is done as follows.

   i. *Generating 3D structures*: To obtain low energy 3D structures of molecules, a rigorous molecular modelling protocol was followed for all compounds. Molecules in the dataset were curated in a downloadable SDF (Standard Data File) format or drawn on a Graphical User Interface (GUI) and adjusted to their ionization states at physiological pH. Molecular geometries were optimized using the Impact tool in Schrödinger Suite 2009 with the OPLS2005

force field[8]. The energy of the molecule was then minimized by the steepest descent algorithm in 100,000 cycles until the energy gradient was less than 0.001 kcal/mol/Å. This was followed by molecular dynamics simulations; the equations of motion were solved by the Verlet integrator. Constraints to hydrogen atoms were defined using the SHAKE algorithm and simulations were performed for 1 ns, with frames captured every 2000 steps. The molecular trajectories were analysed, and the lowest energy structure was identified and further subjected to energy minimization in the same fashion as mentioned above.

ii. *Canonical numbering of atoms*: Since the structures in the dataset are diverse in chemical type and size, it was found necessary to number the structures in a uniform, unambiguous, and unbiased manner. For this, the canonical numbering of structures by the Morgan algorithm[9] was adopted. The algorithm follows a sequential numbering pattern, with atoms iteratively assigned "connectivity values" until each atom is represented by a unique number. The working of the algorithm is illustrated below.

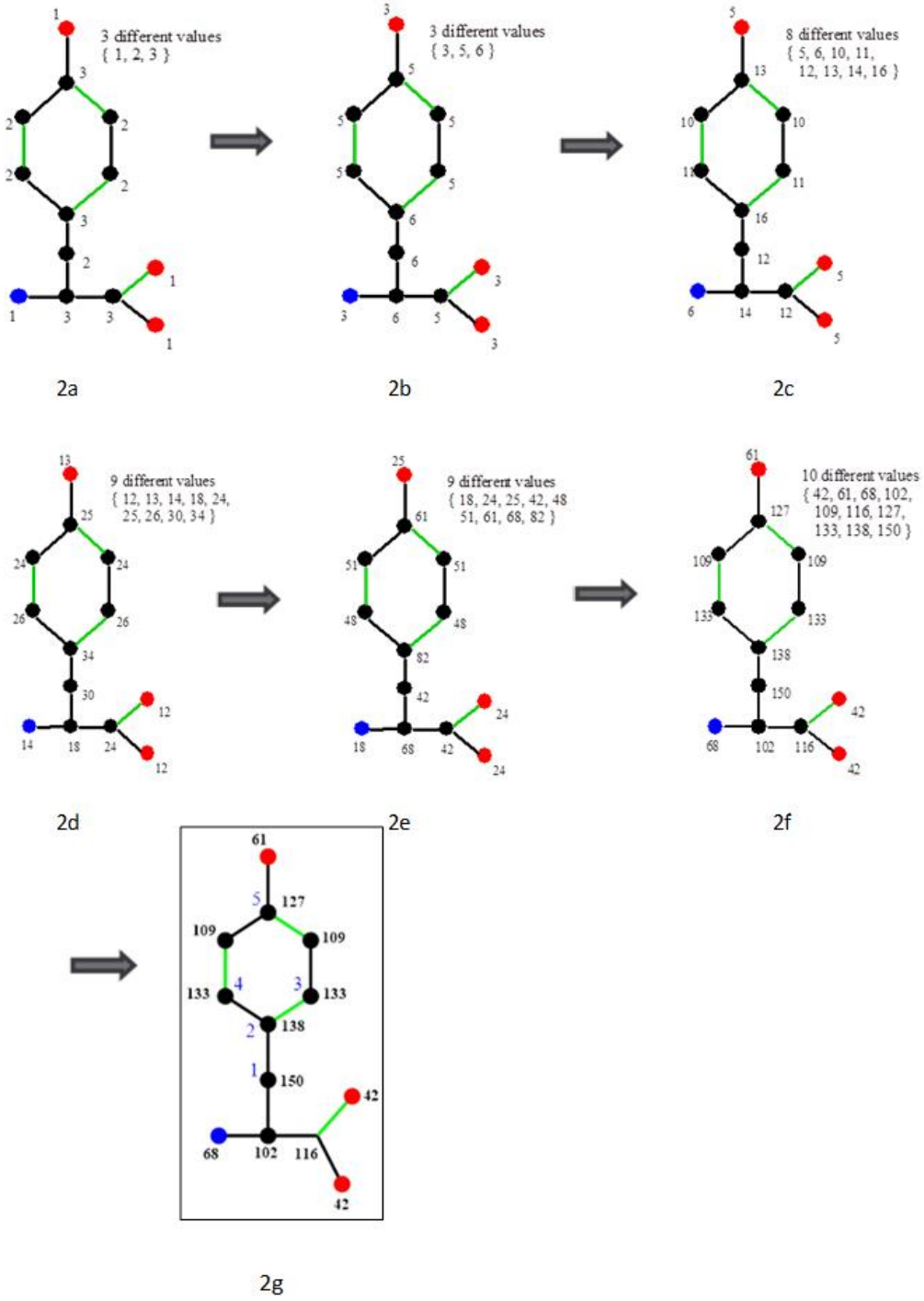


Figure 2: A schematic illustration of the working of the Morgan algorithm. Each node is initially labelled with its degree (a count of the number of connected atoms as shown in 2a). The labels are recalculated by summing values at neighbour nodes. This is repeated until each node has a unique label (2b - 2f). The nodes are subsequently numbered in decreasing order of label values (2g).

iii. Calculation of interatomic distances: Having numbered the atoms, the next step is the calculation of interatomic distances on hydrogen suppressed molecular graphs. A slightly different approach to measure interatomic distances was followed. First, the centroid of the molecule (Y) is computed. For the distance between say atoms numbered 1 and 2, we find the centroid (X) of these two atoms, and the distance between X and Y is a measure of the distance between atoms 1 and 2. We do this for all atom pairs in the molecule. The distances are then populated in a square matrix; the diagonal of the matrix which is the distance of an atom to itself is left empty at this point. Distances were calculated using the Visual Molecular Dynamics package (VMD 1.9.3) package[10]. For an achiral molecule, the upper as well as the lower half of the matrix are identical (Figs. 3 and 4).

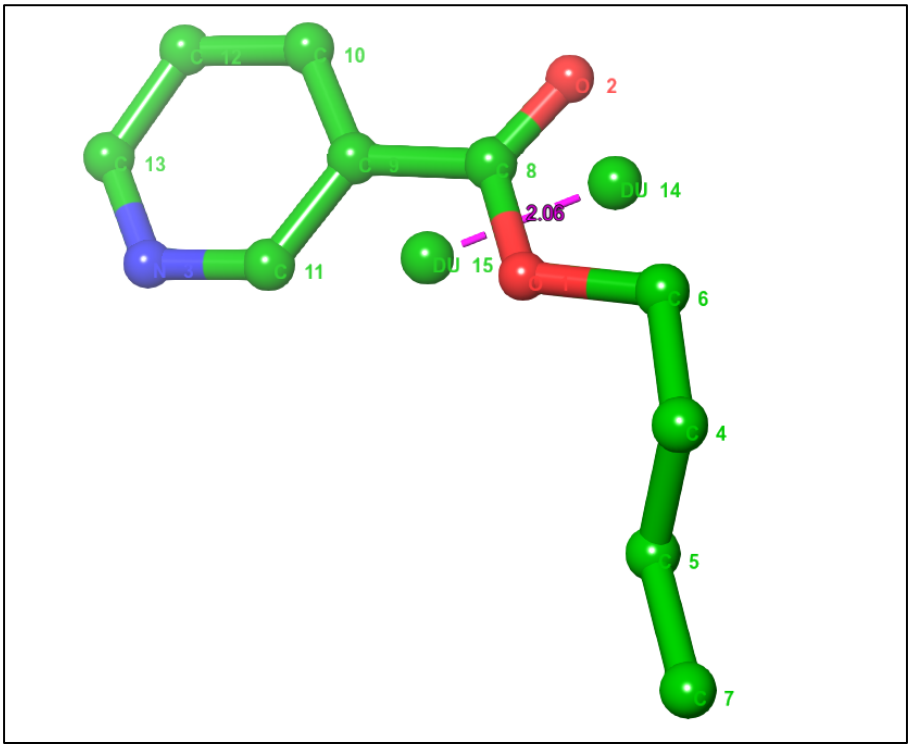


Figure 3: An illustration of the calculation of interatomic distances. DU 14 is the centroid between atom pairs O2 and C6**,** DU 15 is the centroid of the molecule.

	1_C1	10_C2	11_C5	2_C3	3_C4	4_C6	5_O1	6_N1	7_C7	8_C8	9_O2
1_C1		1.377382	1.693203	0.446638	1.203389	0.84946	0.694589	1.501557	1.727719	1.823667	1.04099
10_C2	1.377382		3.496914	2.09719	0.973209	1.401983	2.553251	1.099541	0.331118	0.507444	2.381995
11_C5	1.693203	3.496914		2.395107	1.329819	1.652909	2.802263	1.283526	0.763002	0.742482	2.690926
2_C3	0.446638	2.09719	2.395107		0.907051	0.326174	1.384992	0.795889	1.146782	1.089981	1.612772
3_C4	1.203389	0.973209	1.329819	0.907051		1.189281	0.400032	1.843195	2.367099	2.327209	1.446651
4_C6	0.84946	1.401983	1.652909	0.326174	1.189281		0.973626	2.04891	1.868578	2.194288	0.363014
5_O1	0.694589	2.553251	2.802263	1.384992	0.400032	0.973626		1.071263	0.561366	0.892053	1.679704
6_N1	1.501557	1.099541	1.283526	0.795889	1.843195	2.04891	1.071263		2.533822	2.864726	0.430217
7_C7	1.727719	0.331118	0.763002	1.146782	2.367099	1.868578	0.561366	2.533822		2.980377	1.446185
8_C8	1.823667	0.507444	0.742482	1.089981	2.327209	2.194288	0.892053	2.864726	2.980377		1.05422
9_O2	1.04099	2.381995	2.690926	1.612772	1.446651	0.363014	1.679704	0.430217	1.446185	1.05422	

Figure 4: A sample distance matrix for an achiral molecule. The distances calculated as shown in figure 3 are populated in the off-diagonal cells of a 2D matrix. The diagonal cells are left blank at this point.

iv. *Accounting for chirality*: To incorporate the effect of chirality (molecules with 2 or more chiral centres) on interatomic distances, the distance matrix is suitably modified. This is done as follows: first 3D molecular structures (as described above) are generated for all the stereoisomers, and for each stereoisomer, interatomic distances (as described above) are computed. A distance matrix is then constructed to encompass all the distance attributes of the stereoisomers, thus accounting for the contribution of multiple enantiomeric states. This is done by populating the upper half of the distance matrix with the maximum values of each distance found among the stereoisomers while the lower half is filled with the corresponding minimum values of each distance as seen in the ensemble of stereoisomers, to yield the "max-min" distance matrix for chiral molecules. This is shown in Figure 5.

	1_C6	10_C7	2_C8	3_C2	4_C4	5_C1	6_C3	7_O1	8_O2	9_C5
1_C6		1.673653	2.822772	1.158279	1.608954	0.449017	0.6894	3.338583	2.858703	0.964528
10_C7	1.673653		1.040974	2.916738	2.231847	3.504913	4.176108	0.496863	1.32947	4.76515
2_C8	2.822772	1.040974		1.550736	2.190235	0.940019	0.3076	3.92752	3.526681	0.48429
3_C2	1.158279	2.916738	1.550736		0.523154	1.074307	1.744985	2.093674	1.585935	2.246993
4_C4	1.608954	2.231847	2.190235	0.523154		0.563307	1.021918	2.720153	2.275796	1.631812
5_C1	0.449017	3.504913	0.940019	1.074307	0.563307		2.273933	1.490286	1.137102	2.891959
6_C3	0.6894	4.176108	0.3076	1.744985	1.021918	2.273933		0.833011	0.647242	3.512117
7_O1	3.338583	0.496863	3.92752	2.093674	2.720153	1.490286	0.833011		3.961639	0.407328
8_O2	2.858703	1.32947	3.526681	1.585935	2.275796	1.137102	0.647242	3.961639		0.96906
9_C5	0.964528	4.76515	0.48429	2.246993	1.631812	2.891959	3.512117	0.407328	0.96906	

Figure 5: A "max-min" distance matrix for a chiral molecule. The upper half is filled with maximum values for each distance computed for the stereoisomers while minimum distances are filled in the lower off-diagonal cells. Note that the diagonal cells are empty at this stage.

2. ***Integration of structural and physicochemical information to generate eigenvalues***

Having represented 3D molecular structures as 2D distance matrices, the next phase in the methodology involves merging structural information with physicochemical properties. This is done by incorporating physicochemical property values in the diagonal cells of the max-min distance matrix, followed by mathematically transforming the "property integrated distance matrix" to generate eigenvalues. The steps are enumerated below.

i. *Physicochemical property calculation*: First, physicochemical properties encoding characteristic features are computed for each molecule. To capture intricacies in properties at the atomic level, atomic properties like atomic partial charge, atomic polarizability, atomic orbital electronegativity, atomic solvent accessible surface area, and atomic contributions to van der Waals' surface area and refractivity were calculated. Molecular properties included atom and group counts, atom types and hybridisation states, electrotopological states[11,12,13], constitutional properties, shape indices[14], volume[15], lipophilicity[16], polar surface area[17], and

topological charges[18]. Atomic properties were computed using the Marvin tool of the ChemAxon Calculator plugin, version 18.8.0 Molecular properties were calculated using the software PaDEL[19].

ii. *Property pool refinement*: An initial 1876 properties (atomic plus molecular) were calculated for each molecule and these were filtered by objective feature selection. Standard deviation cut-offs were initially used to remove redundant properties. Subsequently, properties for which 80% of the values were identical in all compounds were discarded; this was followed by removing properties if their pairwise correlation coefficient was high. Scaling of the filtered physicochemical properties was avoided as it was observed that unscaled values are better suited to the EVANS methodology.

iii. *Construction of PD Matrices (Physicochemical Property Integrated Distance Matrices)*: The inherent physicochemical properties of the molecules are integrated with its spatial topology by populating the diagonal cells of the max-min distance matrices with a molecular/atomic property; this step yields as many PD matrices for a molecule as the total number of molecular/atomic properties calculated. A variance-covariance matrix is then computed for each PD matrix to ensure the spread of properties along the topology of the molecule (shown in figures 6 and 7).

	1_C1	10_C7	2_C2	3_O1	4_C3	5_O2	6_C4	7_C5	8_C6	9_O3
1_C1	-0.6412	2.632713	0.517936	0.740355	1.297505	2.172332	0.807617	2.541439	1.127152	1.228274
10_C7	2.179934	-0.6412	1.132023	1.685039	1.927802	3.081833	1.318916	3.08159	1.768669	1.415765
2_C2	0.208107	0.982014	-0.6412	1.423674	0.989788	0.762608	2.404918	0.966161	2.611027	2.542567
3_O1	0.534321	1.460599	1.229112	-0.6412	0.408284	1.178138	1.770845	1.35449	2.053104	2.025688
4_C3	1.087353	1.48979	0.899688	0.304826	-0.6412	1.542883	1.356517	2.035802	1.555481	1.546
5_O2	1.946622	2.710562	0.361298	0.930643	1.22277	-0.6412	0.976851	2.691331	1.38449	1.142257
6_C4	0.501114	0.527399	2.225091	1.526183	1.187669	0.235493	-0.6412	1.32555	2.762071	3.167262
7_C5	2.331093	2.514431	0.750538	1.11836	1.801097	2.439583	0.243844	-0.6412	1.52294	1.705063
8_C6	0.655797	0.451345	2.320805	1.658733	1.183	0.572601	2.551468	0.205282	-0.6412	2.829608
9_O3	0.964492	0.893328	2.223696	1.599559	1.404178	0.535224	2.798123	0.813007	2.457813	-0.6412

Figure 6: A PD matrix: Physicochemical property (ALogP) has been filled in the diagonal cells, generating one such PD matrix. Multiple PD matrices can be constructed by replacing the physicochemical property in the diagonal.

	1_C1	10_C7	2_C2	3_O1	4_C3	5_O2	6_C4	7_C5	8_C6	9_O3
1_C1	0.885388	-0.0711	0.0121	0.132431	0.207905	0.181796	-0.16361	-0.05396	-0.01909	-0.21704
10_C7	-0.0711	1.183811	-0.45891	-0.37447	-0.10967	-0.18889	-0.31879	-0.09331	-0.10647	-0.24916
2_C2	0.0121	-0.45891	0.901672	0.160205	0.063268	-0.16517	0.005186	-0.27017	-0.18801	-0.03548
3_O1	0.132431	-0.37447	0.160205	0.552234	0.353335	-0.04355	0.041679	-0.1353	-0.04573	0.001153
4_C3	0.207905	-0.10967	0.063268	0.353335	0.539508	0.16863	-0.1117	-0.10654	-0.02129	-0.10378
5_O2	0.181796	-0.18889	-0.16517	-0.04355	0.16863	1.276338	-0.17036	0.110149	-0.06513	-0.10631
6_C4	-0.16361	-0.31879	0.005186	0.041679	-0.1117	-0.17036	1.161407	-0.16761	-0.23288	-0.3847
7_C5	-0.05396	-0.09331	-0.27017	-0.1353	-0.10654	0.110149	-0.16761	1.369189	0.133383	-0.25648
8_C6	-0.01909	-0.10647	-0.18801	-0.04573	-0.02129	-0.06513	-0.23288	0.133383	0.956325	-0.17015
9_O3	-0.21704	-0.24916	-0.03548	0.001153	-0.10378	-0.10631	-0.3847	-0.25648	-0.17015	1.151563

Figure 7: A variance-covariance matrix.

iv. *Calculation of eigenvalues*: Finally, the covariance matrix is diagonalized to give the eigenvalues and eigenvectors (computed using the R software environment for statistical computing, version 4.0.3)[20]. The first three eigenvalues (possessing the maximum information content) for each PD matrix are taken forward as independent variables (descriptors) for building the models.

3. *Model building and validation*

The final step in the EVANS methodology involves correlating the biological endpoints (the dependent variables, Y) with the molecular eigenvalue descriptors (the independent variables, X) using chemometric methods, to derive a meaningful correlation. Since PK properties are known to exhibit nonlinearity, we employed nonlinear machine learning algorithms (random forests and support vector machine) along with the traditional multiple linear regression for building predictive QSPKR models. This served an additional objective of investigating the chemometric methods best suited to the methodology.

i. *Division into training/test sets:* After some deliberation, we opted against the implementation of algorithms for training and test set selection since a benchmarking study by Martin et al., 2012[21] found no significant difference in model quality on using random division versus algorithms such as sphere exclusion, Kennard stone and minimal test set similarity. Rational selection of training and test sets was thus carried out by sorting molecules according to biological activity and assigning the "n^{th}" molecule to the test set. The complete training and test set data along with computed eigenvalues are provided as csv files in the Supplementary Material.

ii. *Feature Selection and Regression Analyses:* Models were built by correlating the three PK endpoints (VD_{ss}, CL, and $t_{1/2}$) with the eigenvalue descriptors using suitable variable selection and chemometric methods. Eklund et al., 2012[22] investigated the performance of a range of feature selection methods for QSAR studies, and in accordance with their findings, the Multi Adaptive Regression Splines (MARS) approach was used for objective feature selection. All QSPKR models were built using the R program for statistical computing, version 4.0.3[20]. Models were built with the following chemometric methods.

- Multiple linear regression fitness evaluator
- Random forest algorithm with the "randomForest" package in R. The number of trees was varied from 200-500, and the number of predictors sampled for splitting at each node was kept at the default value (p/3 where p is the number of independent variables).
- Support vector regression models were built with the "e1071" package using linear and radial basis function kernels. Tuneable parameters (cost and sigma) were optimized before model building.

Internal validation metrics were computed using the "caret" package in R, and the Xternal Validation Plus program (version 1.2)[23] was used for evaluation of predictive performance on the test set. The model which passed all statistical diagnostics was chosen as the optimum model.

RESULTS

Analysis of features used for model building

In a complex modelling procedure, model interpretation is always a daunting task. While we acknowledge the pitfalls in attempting to ascribe physical meaning to mutually orthogonal eigenvalues, we have analysed the important descriptors used for the regression analyses to better understand the nature of the models. Identified using the MARS algorithm, the important features are given in Tables 2 to 4. It is well established that charge, lipophilicity, and polarity play an important role in determining the pharmacokinetic profile of drug molecules[24]. It is evident that our variable selection routine captures all these features; eigenvalues encoding charge (*atmioncharge_EigV2, atmioncharge_EigV3, DELS2_EigV3, MAXDN_EigV2, MAXDP_EigV3*), lipophilicity (*LipoaffinityIndex_EigV3, XLogP_EigV2*), and polarity information (*TopoPSA_EigV1, TopoPSA_EigV2, bpol_EigV1, bpol_EigV2*) account for 14 of the total 29 features identified for VD_{ss}, CL and $t_{1/2}$ models. Other features include electro-topological state variables (*SHBa_EigV2, gmin_EigV1*), bond counts, (*nBondsS_EigV2*), atomic contributions to refractivity (*atmrefract_EigV1, atmrefract_EigV3*), and surface area (*atmmsavdw_EigV1, atmmsavdw_EigV2*). Furthermore, we validated our models by comparing the features used to build the EVANS models with some published QSPKR models in the literature. Despite the use of different datasets and various tools for feature calculation and selection, the nature of the descriptors used in QSPKR model building appears to be the same, with descriptors encoding lipophilicity, polarity, charge, E-state, and surface area frequently appearing in the published models[25,26,27,28].

Table 2: A brief description of the features used to build the most statistically significant QSPKR models for VD_{ss}

Term	Description
SHBa_EigV2	Second eigenvalue of the sum of E-States for (strong) hydrogen bond acceptors
nBondsS_EigV2	Second eigenvalue of the total number of single bonds (including bonds to hydrogens, excluding aromatic bonds)
atmrefract_EigV3	Third eigenvalue of the atomic refractivity
TopoPSA_EigV2	Second eigenvalue of the topological polar surface area
MAXDN_EigV2	Second eigenvalue of the maximum negative intrinsic state difference in the molecule (related to the nucleophilicity of the molecule)
atmioncharge_EigV2	Second eigenvalue of the atom-wise ion charge

atmmsavdw_EigV2	Second eigenvalue of the atom-wise molecular surface area
LipoaffinityIndex_EigV3	Third eigenvalue of the Lipoaffinity Index
gmin_EigV1	First eigenvalue of the minimum E-State
TopoPSA_EigV1	First eigenvalue of the topological polar surface area

Table 3: A brief description of the features used to build the most statistically significant QSPKR models for CL

Term	**Description**
TopoPSA_EigV1	First eigenvalue of the topological polar surface area
bpol_EigV1	First eigenvalue of the sum of the absolute value of the difference between atomic polarizabilities of all bonded atoms in the molecule
XLogP_EigV2	Second eigenvalue of XlogP
atmmsavdw_EigV2	Second eigenvalue of the atom-wise molecular surface area
nHBa_EigV2	Second eigenvalue of the number of H-bond acceptors
atmioncharge_EigV3	Third eigenvalue of the atom-wise ion charge
TopoPSA_EigV2	Second eigenvalue of the topological polar surface area
atmmsavdw_EigV1	First eigenvalue of the atom wise molecular surface area
ETA_Eta_F_EigV3	Third eigenvalue of the functionality index EtaF

Table 4: A brief description of the features used to build the most statistically significant QSPkR models for $t_{1/2}$

Term	**Description**
TopoPSA_EigV2	Second eigenvalue of the topological polar surface area
DELS2_EigV3	Third eigenvalue of the sum of all atoms intrinsic state differences (a measure of total charge transfer in the molecule)
nHBa_EigV3	Third eigenvalue of the number of H-bond acceptors
MAXDP_EigV3	Third eigenvalue of the maximum positive intrinsic state difference in the molecule (related to the electrophilicity of the molecule)
ETA_Eta_R_EigV2	Second eigenvalue Composite index Eta for reference alkane
nBondsS3_EigV3	Third eigenvalue of the total number of single bonds (excluding bonds to hydrogens and aromatic bonds)
atmmsavdw_EigV2	Second eigenvalue of the atomwise molecular surface area
atmrefract_EigV1	First eigenvalue of the atomic refractivity
bpol_EigV2	Second eigenvalue of the sum of the absolute value of the difference between atomic polarizabilities of all bonded atoms in the molecule
atmrefract_EigV3	Third eigenvalue of the atomic refractivity

Evaluating the performance of EVANS across different chemometric methods

Four chemometric methods: MLR, RF, linear SVM, and nonlinear SVM were employed for model building using eigenvalues (descriptors) selected by the MARS algorithm. EVANS models were built

on the training set and evaluated using the standard r^2 metric which represents the proportion of variance that is explained by the model, and on error-based metrics such as root mean squared error (RMSE) and mean absolute error (MAE). Models were validated internally using leave-one-out (q^2_{loo})and 10-fold cross-validation [$q^2_{(10-fold)}$]. The predictive performance of the model was evaluated on an independent test set and measured using the r^2_{pred}, $RMSE_{test}$, MAE_{test}, and r^2_m metrics.

The models built for VD_{ss} show good correlations on the training set; all four algorithms resulted in models with $r^2 \geq 0.60$, with the nonlinear SVM model having r^2 of 0.69. However, we argue that the best VD_{ss} model is the one from the RF algorithm since it is assuredly robust (q^2_{loo} = 0.58) with the best predictive performance (r^2_{pred} = 0.59). CL models have relatively lower r^2 values, with the three algorithms resulting in models with consistently explained variance (r^2 = 0.38). For this endpoint, the best internal validation metrics (q^2_{loo} = 0.35) and predictive ability (r^2_{pred} = 0.29) arose from linear SVM. Models built for $t_{1/2}$ show comparatively larger variations with average performance in both training and test set; this was somewhat expected since $t_{1/2}$ is a secondary PK parameter that derives its value from VD_{ss} and CL. The best training set correlations were obtained using nonlinear SVM (r^2 = 0.51) with r^2_{pred} = 0.21. The complete validation metrics for all models are given in **Table 5**.

Table 5: Statistical parameters and equations for EVANS models obtained using four chemometric methods

Algorithm	r^2_{train}	RMSE	MAE	q^2_{loo}	$q^2_{(10-fold)}$	r^2_{pred}	$RMSE_{ext}$	MAE_{test}	Average r^2_m
VD_{ss}									
MLR	0.61	0.30	0.25	0.56	0.59	0.42	0.32	0.28	0.21
RF	0.60	0.30	0.24	0.59	0.58	0.59	0.29	0.25	0.42
Linear SVM	0.64	0.29	0.23	0.62	0.62	0.41	0.33	0.28	0.25
Nonlinear SVM	0.69	0.26	0.20	0.58	0.59	0.53	0.31	0.27	0.33
CL									
MLR	0.38	0.28	0.23	0.33	0.34	0.25	0.29	0.24	0.11
RF	0.28	0.29	0.24	0.29	0.30	0.24	0.29	0.25	0.02
Linear SVM	0.38	0.28	0.23	0.32	0.35	0.29	0.29	0.24	0.12
Nonlinear SVM	0.38	0.27	0.22	0.28	0.28	0.25	0.28	0.23	0.06
$t_{1/2}$									
MLR	0.38	0.29	0.25	0.33	0.35	0.28	0.27	0.22	0.09
RF	0.41	0.28	0.24	0.41	0.46	0.1	0.29	0.25	0.01
Linear SVM	0.41	0.28	0.24	0.36	0.38	0.3	0.27	0.22	0.11
Nonlinear SVM	0.51	0.25	0.21	0.34	0.37	0.21	0.27	0.23	0.05

DISCUSSION

QSPR modelling has been a cornerstone of *in silico* drug discovery and development for several decades. The traditional Hansch approach has now been supplemented with 3D, 4D, 5D, and even 6D QSAR methods[2]. However, multidimensional QSAR/QSPR suffers from two major drawbacks: the need to perform an accurate alignment procedure and the difficulties in dealing with drugs having multiple chiral centres. We have attempted to circumvent these problems by developing a QSPR methodology titled "Eigenvalue Analysis (EVANS)" that incorporates 3D structural information without the need to perform molecular alignment. We do so by projecting interatomic distances calculated on low energy structures onto the off-diagonal cells of a "distance" matrix. To account for chirality, we elucidate all enantiomeric states of molecules with more than one chiral centre and compute distances on all low energy structures. The maximum and minimum values of these computed distances then populate the upper and lower cells respectively of the distance matrix, thus generating a "max-min" distance matrix. This, in some way, accounts for the ensemble of structures that are likely to interact with the target receptor. The max-min distance matrices are then integrated with molecular and atomic physicochemical properties in the diagonal cells, resulting in property integrated distance matrices (PD matrices) that are distinct for each property and molecule. For example, 100 physicochemical properties for a dataset of 200 molecules would generate 100 * 200 = 20,000 PD matrices. These PD matrices are mathematically transformed to generate mutually orthogonal eigenvalues for each property and each molecule. The eigenvalues are thus a hybrid of 3D structure and physicochemical properties and form an unbiased numerical representation of molecular structure and property, which is the essence of the EVANS methodology. The final step involves the eigenvalue descriptors as independent variables in QSPR modelling.

In a previous study[7], we have tested the EVANS methodology on pharmacodynamic datasets with encouraging results. This study focuses on the application of EVANS to build predictive QSPKR models using clinically derived PK data curated by Obach et al.[29] The PK parameters modelled were VD_{ss}, CL, and $t_{1/2}$, and models were built using various chemometric methods like MLR, RF, and SVM with linear and radial basis function kernels. The initial models built on the entire dataset of 474 molecules were refined, and the models for VD_{ss} display the best training set correlation and predictive ability (r^2 ranges from 0.60 to 0.69 and r^2_{pred} varies from 0.41 to 0.59) across the four chemometric methods. The r^2_{train} *and* r^2_{pred} values for CL and $t_{1/2}$ are slightly lower (r^2_{train} 0.28 to 0.38 and r^2_{pred} 0.24 to 0.29 for CL and r^2_{train} 0.38 to 0.51 and r^2_{pred} 0.10 to 0.30 for $t_{1/2}$ respectively) than the corresponding values for the models for VD_{ss}

To gauge the effectiveness of the EVANS formalism, we compare the validation metrics with the QSPKR models reported in the literature. It is known that the VD_{ss} of a drug is dictated to a large extent by its acidity or basicity, with the ionization state influencing plasma protein binding and distribution

in extracellular space[30]. This theme has been explored by Zhivkova et al., 2011[31] and Zhivkova et al., 2015[25] who have predicted VD_{ss} by building separate QSPKR models for acids and bases. They report r^2_{train} of 0.66 and r^2_{pred} of 0.01 for 132 acidic drugs, whereas the model for 216 bases had r^2_{train} = 0.66 and r^2_{pred} = 0.59. A similar approach was followed by Simeon et al., 2019[32], who built global models on a large dataset of 1442 chemicals for human VD_{ss} using RF, Artificial Neural Nets (ANNs), and Partial Least Squares (PLS). Their consensus models had an r^2_{train} of 0.64. The dataset was then divided into clusters based on chemical class, and the consensus model was employed to make predictions for each class. In addition to single models, Lombardo et al., 2016[27] propose a two-tier modelling approach wherein a tier-one classification model is combined with a tier-two range specific regression model. Although they did not find a significant improvement over the single model, the idea is an intriguing one and may find more applicability in future work, especially in the case of diverse datasets with a large spread of VD_{ss}. The EVANS models for CL are also comparable with those reported in the literature, e.g. Dave et al., 2015[26] report q^2_{loo} of 0.14 on a single model built on a dataset of 382 drug-like compounds. They hypothesize that one model may struggle to accurately predict compounds that vary in their ion-status, affinity for transporters, and mechanism of elimination, as all these factors play a key role in determining the clearance of a molecule. Chen et al., 2020[33] results are in agreement with Dave et al., 2015[26]; their global models have q^2_{loo} of 0.14 and 0.20 with MLR and RF algorithms respectively, which led them to adopt an ionization-state and elimination route based approach for clustering the data. While we were unable to find too many studies focused on the prediction of $t_{1/2}$, we note that Arnot et al., 2014[34] and Lu et al., 2016[35] built molecules on a dataset of 1104 organic chemicals with promising results. There was one study that stood out, Wang et al., 2019[28] who have reported four human PK parameters including VD_{ss}, CL, and $t_{1/2}$ on the dataset curated by Lombardo et al., 2018[36]. They present r^2_{train} of 0.95 and r^2_{pred} of 0.87 for VD_{ss}, r^2_{train} and r^2_{pred} both 0.88 for CL, and r^2_{train} of 0.88 and r^2_{pred} of 0.83 for $t_{1/2}$. A comparison of the EVANS models with models published in the literature is given in **Table 6**.

Table 6: A comparison of the EVANS QSPKR models with some published models in the literature

Reference	**No. of molecules**	**Chemometric method**	r^2_{train}	q^2_{loo}	**RMSE$_{train}$**	r^2_{pred}
VD_{ss}						
EVANS model	328	RF	0.60	0.59	0.30	0.59
Zhivkova et al., 2011[31]	132	Stepwise regression	0.66	0.58	-	0.01
Zhivkova et al., 2015[25]	216	MLR	0.66	0.61	-	0.59

Simeon et al., 2019[32]	1441	Consensus (PLS, RF and ANN)	0.64	-	0.41	0.57
Wang et al., 2019[28]	1270	SVM	0.95	0.76	0.14	0.87
CL						
EVANS model	328	Linear SVM	0.38	0.28	0.32	0.29
Dave et al., 2015[26]	332	Stepwise regression	-	0.14	-	-
Chen et al., 2020[33]	636	MLR	0.21	0.13	1.77	0.17
Chen et al., 2020[33]	636	RF	0.36	0.20	1.52	0.20
Wang et al., 2019[28]	1270	RF	0.88	0.78	0.24	0.83
$t_{1/2}$						
EVANS model	328	Linear SVM	0.41	0.36	0.28	0.3
Arnot et al., 2014[34]	470	Iterative Fragment Selection	0.89	-	0.47	0.73
Lu et al., 2016[35]	1105	Gradient Boosting Machine	0.96	-	0.28	0.82
Wang et al., 2019[28]	1270	RF	0.88	0.73	0.21	0.83

From the studies mentioned above, it is clear that the accepted strategy for PK modelling is to build "fit for purpose" or local models, with data assigned into clusters based on the nature and quality of the dataset and objectives of the study. While this may result in a lower domain of applicability, the resulting models may be more accurate owing to a lack of confusing structure-activity relationships that arise due to structurally and biologically diverse molecules in the same training set. The consensus appears to be that a single or global QSPKR model may not be able to capture the complexities of the multiple PK processes that govern the pharmacokinetics of drug-like molecules. With this in mind, we feel the EVANS models for VD_{ss}, CL, and $t_{1/2}$ built with the hybrid eigenvalues stand in good stead. We are currently working towards a more targeted approach that will hopefully yield better results.

CONCLUSION

This work focuses on the extensibility of our earlier reported QSPR formalism entitled "EigenValue ANalySis (EVANS)" to build predictive models for human intravenous PK data. The EVANS

methodology uses 3D structural information with due consideration of all enantiomeric states for chiral molecules, hybridized with molecular and atomic physicochemical properties to generate eigenvalues, that are used as independent variables in QSPR analyses. EVANS has previously been tested on pharmacodynamic datasets with promising results. In this paper, we have built QSPKR models using the EVANS formalism for three critical PK parameters: VD_{ss}, CL, and $t_{1/2}$. Models were built using the traditional MLR approach, along with machine learning algorithms such as RF and SVM with both linear and nonlinear kernels, and the EVANS models for VD_{ss} show especially encouraging results. An analysis of the QSPKR models reported in the literature illustrates the complexities of *in silico* PK modelling. In comparison, the global EVANS QSPKR models which have been built on a large and diverse dataset stand in good stead. Efforts are currently being directed at expanding the methodology to model toxicity endpoints and to predict the permeability of chemicals across biological membranes. We hope that with further refinement, EVANS will be adopted into a useful first-line prediction tool to prioritize compounds in drug discovery and development.

ACKNOWLEDGEMENT

The authors are thankful to the Department of Biotechnology, Government of India for funding manpower and consumables for this research (Order No: BT/PR13600/BID/7/545/2015 dated 29/12/2016).

REFERENCES

(1) Roy, K.; Kar, S.; Das, R. N. *Understanding the Basics of QSAR for Applications in Pharmaceutical Sciences and Risk Assessment*; 2015. https://doi.org/10.1016/C2014-0-00286-9.

(2) Verma, J.; Khedkar, V.; Coutinho, E. 3D-QSAR in Drug Design - A Review. *Curr. Top. Med. Chem.* **2010**, *10* (1), 95–115. https://doi.org/10.2174/156802610790232260.

(3) Eichelbaum, M.; Gross, A. S. *Stereochemical Aspects of Drug Action and Disposition*; 1996; Vol. 28. https://doi.org/10.1016/s0065-2490(96)80003-7.

(4) Saghir, S. A.; Rais, A. A. Pharmacokinetics. **2018**, 1–9. https://doi.org/10.1016/B978-0-12-801238-3.62154-2.

(5) Kenakin, T. Pharmacokinetics. In *A Pharmacology Primer*; Elsevier Inc., 2019; pp 245–293. https://doi.org/10.1016/B978-0-12-813957-8.00009-6.

(6) Mansoor, A.; Mahabadi, N. Volume of Distribution https://www.ncbi.nlm.nih.gov/books/NBK545280/ (accessed Dec 10, 2020).

(7) Joseph, B.; Gomatam, A. N.; Shaikh, M. A. S.; Khedkar, V.; Martis, E. A. F.; Coutinho, E. C. EigenValue ANalySis (EVANS) – A Tool to Address Pharmacodynamic, Pharmacokinetic and Toxicity Issues. *Int. J. Quant. Struct. Relationships* **2019**, *4*(3), 118–136. https://doi.org/10.4018/ijqspr.2019070105.

(8) Jorgensen, W. L.; Maxwell, D. S.; Tirado-Rives, J. Development and Testing of the OPLS All-Atom Force Field on Conformational Energetics and Properties of Organic Liquids. *J. Am. Chem. Soc.* **1996**, *118* (45), 11225–11236. https://doi.org/10.1021/ja9621760.

(9) Morgan, H. L. The Generation of a Unique Machine Description for Chemical Structures—A Technique Developed at Chemical Abstracts Service. *J. Chem. Doc.* **1965**, *5* (2), 107–113. https://doi.org/10.1021/c160017a018.

(10) Humphrey, W.; Dalke, A.; Schulten, K. VMD: Visual Molecular Dynamics. *J. Mol. Graph.* **1996**, *14* (1), 33–38. https://doi.org/10.1016/0263-7855(96)00018-5.

(11) Gramatica, P.; Corradi, M.; Consonni, V. Modelling and Prediction of Soil Sorption Coefficients of Non-Ionic Organic Pesticides by Molecular Descriptors. *Chemosphere* **2000**, *41* (5), 763–777. https://doi.org/10.1016/S0045-6535(99)00463-4.

(12) Hall, L. H.; Kier, L. B. Electrotopological State Indices for Atom Types: A Novel Combination of Electronic, Topological, and Valence State Information. *J. Chem. Inf. Comput. Sci.* **1995**, *35* (6), 1039–1045. https://doi.org/10.1021/ci00028a014.

(13) Liu, R.; Sun, H.; So, S. S. Development of Quantitative Structure-Property Relationship Models for Early ADME Evaluation in Drug Discovery. 2. Blood-Brain Barrier Penetration. *J. Chem. Inf. Comput. Sci.* **2001**, *41* (6), 1623–1632. https://doi.org/10.1021/ci010290i.

(14) Hall, L. H.; Kier, L. B. The Molecular Connectivity Chi Indexes and Kappa Shape Indexes in Structure-Property Modeling. In *Reviews in Computational Chemistry*; Lipkowitz, K., Boyd, D.,

Eds.; Wiley-VCH, Inc., 1991; Vol. 2, pp 367–422. https://doi.org/10.1002/9780470125793.ch9.

(15) Abraham, M. H.; McGowan, J. C. The Use of Characteristic Volumes to Measure Cavity Terms in Reversed-Phase Liquid Chromatography. *Chromatographia* **1987**, *23* (4), 243–246. https://doi.org/10.1007/BF02311772.

(16) Mannhold, R.; Poda, G.; Ostermann, C.; Tetko, I. Calculation of Molecular Lipophilicity: State-of-the-Art and Comparison of Log P Methods on More Than 96,000 Compounds. *J. Pharm. Sci.* **2009**, *98* (3), 861–893. https://doi.org/10.1002/jps.

(17) Ertl, P.; Rohde, B.; Selzer, P. Fast Calculation of Molecular Polar Surface Area as a Sum of Fragment-Based Contributions and Its Application to the Prediction of Drug Transport Properties. *J. Med. Chem.* **2000**, *43* (20), 3714–3717. https://doi.org/10.1021/jm000942e.

(18) Todeschini, R.; Consonni, V. *Molecular Descriptors for Chemoinformatics*; 2010; Vol. 2. https://doi.org/10.1002/9783527628766.

(19) Yap, C. W. PaDEL-Descriptor: An Open Source Software to Calculate Molecular Descriptors and Fingerprints. *J. Comput. Chem.* **2011**, *32* (7), 1466–1474. https://doi.org/10.1002/jcc.21707.

(20) Team, R. C. R: A Language and Environment for Statistical Computing. 2013.

(21) Martin, T. M.; Harten, P.; Young, D. M.; Muratov, E. N.; Golbraikh, A.; Zhu, H.; Tropsha, A. Does Rational Selection of Training and Test Sets Improve the Outcome of QSAR Modeling? *J. Chem. Inf. Model.* **2012**, *52* (10), 2570–2578. https://doi.org/10.1021/ci300338w.

(22) Eklund, M.; Norinder, U.; Boyer, S.; Carlsson, L. Benchmarking Variable Selection in QSAR. *Mol. Inform.* **2012**, *31* (2), 173–179. https://doi.org/10.1002/minf.201100142.

(23) Roy, K.; Das, R. N.; Ambure, P.; Aher, R. B. Be Aware of Error Measures. Further Studies on Validation of Predictive QSAR Models. *Chemom. Intell. Lab. Syst.* **2016**, *152*, 18–33. https://doi.org/10.1016/j.chemolab.2016.01.008.

(24) Jambhekar, S. Physicochemical and Biopharmaceutical Properties of Drug Substances and Pharmacokinetics. In *Foye's Principles of Medicinal Chemistry*; Lemke, T., Williams, D., Roche, V., Zito, W., Eds.; Lippincott Williams & Wilkins, 2013.

(25) Zhivkova, Z.; Mandova, T.; Doytchinova, I. Quantitative Structure – Pharmacokinetics Relationships Analysis of Basic Drugs: Volume of Distribution. *J. Pharm. Pharm. Sci.* **2015**, *18* (3), 515–527. https://doi.org/10.18433/j3xc7s.

(26) Dave, R. A.; Morris, M. E. Quantitative Structure-Pharmacokinetic Relationships for the Prediction of Renal Clearance in Humans. *Drug Metab. Dispos.* **2015**, *43* (1), 73–81. https://doi.org/10.1124/dmd.114.059857.

(27) Lombardo, F.; Jing, Y. In Silico Prediction of Volume of Distribution in Humans. Extensive Data Set and the Exploration of Linear and Nonlinear Methods Coupled with Molecular Interaction Fields Descriptors. *J. Chem. Inf. Model.* **2016**, *56* (10), 2042–2052. https://doi.org/10.1021/acs.jcim.6b00044.

(28) Wang, Y.; Liu, H.; Fan, Y.; Chen, X.; Yang, Y.; Zhu, L.; Zhao, J.; Chen, Y.; Zhang, Y. In Silico

Prediction of Human Intravenous Pharmacokinetic Parameters with Improved Accuracy. *J. Chem. Inf. Model.* **2019**, *59* (9), 3968–3980. https://doi.org/10.1021/acs.jcim.9b00300.

(29) Obach, R. S.; Lombardo, F.; Waters, N. J. Trend Analysis of a Database of Intravenous Pharmacokinetic Parameters in Humans for 670 Drug Compounds. *Drug Metab. Dispos.* **2008**, *36* (7), 1385–1405. https://doi.org/10.1124/dmd.108.020479.

(30) Xu, C.; Mager, D. E. Quantitative Structure -- Pharmacokinetic Relationships. *Expert Opin. Drug Metab. Toxicol.* **2011**, *7* (1), 63–77. https://doi.org/10.1517/17425255.2011.537257.

(31) Zhivkova, Z.; Doytchinova, I. Prediction of Steady-State Volume of Distribution of Acidic Drugs by Quantitative Structure–Pharmacokinetics Relationships. *J. Pharm. Sci.* **2012**, *101* (3), 1253–1266.

(32) Simeon, S.; Montanari, D.; Gleeson, M. P. Investigation of Factors Affecting the Performance of in Silico Volume Distribution QSAR Models for Human, Rat, Mouse, Dog & Monkey. *Mol. Inform.* **2019**, *38* (10), 1–12. https://doi.org/10.1002/minf.201900059.

(33) Chen, J.; Yang, H.; Zhu, L.; Wu, Z.; Li, W.; Tang, Y.; Liu, G. In Silico Prediction of Human Renal Clearance of Compounds Using Quantitative Structure-Pharmacokinetic Relationship Models. *Chem. Res. Toxicol.* **2020**, *33* (2), 640–650. https://doi.org/10.1021/acs.chemrestox.9b00447.

(34) Arnot, J. A.; Brown, T. N.; Wania, F. Estimating Screening-Level Organic Chemical Half-Lives in Humans. *Environ. Sci. Technol.* **2014**, *48* (1), 723–730. https://doi.org/10.1021/es4029414.

(35) Lu, J.; Lu, D.; Zhang, X.; Bi, Y.; Cheng, K.; Zheng, M.; Luo, X. Estimation of Elimination Half-Lives of Organic Chemicals in Humans Using Gradient Boosting Machine. *Biochim. Biophys. Acta - Gen. Subj.* **2016**, *1860* (11), 2664–2671. https://doi.org/10.1016/j.bbagen.2016.05.019.

(36) Lombardo, F.; Berellini, G.; Obach, R. S. Trend Analysis of a Database of Intravenous Pharmacokinetic Parameters in Humans for 1352 Drug Compounds. *Drug Metab. Dispos.* **2018**, *46* (11), 1466–1477. https://doi.org/10.1124/dmd.118.082966.