CrossMark

ORIGINAL ARTICLE

Application of two-dimensional binary fingerprinting methods for the design of selective Tankyrase I inhibitors

B. S. Muddukrishna[1] · Vasudev Pai[2] · Richard Lobo[2] · Aravinda Pai[3]

Abstract In the present study, five important binary fingerprinting techniques were used to model novel flavones for the selective inhibition of Tankyrase I. From the fingerprints used: the fingerprint *atom pairs* resulted in a statistically significant 2D QSAR model using a kernel-based partial least square regression method. This model indicates that the presence of electron-donating groups positively contributes to activity, whereas the presence of electron withdrawing groups negatively contributes to activity. This model could be used to develop more potent as well as selective analogues for the inhibition of Tankyrase I.

✉ Aravinda Pai
aravind.pai@manipal.edu

[1] Department of Pharmaceutical Quality Assurance, Manipal College of Pharmaceutical Sciences (MCOPS), Manipal University, Manipal, Karnataka, India

[2] Department of Pharmacognosy, Manipal College of Pharmaceutical Sciences (MCOPS), Manipal University, Manipal, Karnataka, India

[3] Department of Pharmaceutical Chemistry, Manipal College of Pharmaceutical Sciences (MCOPS), Manipal University, Manipal, Karnataka 576 104, India

Graphical Abstract Schematic representation of 2D QSAR work flow
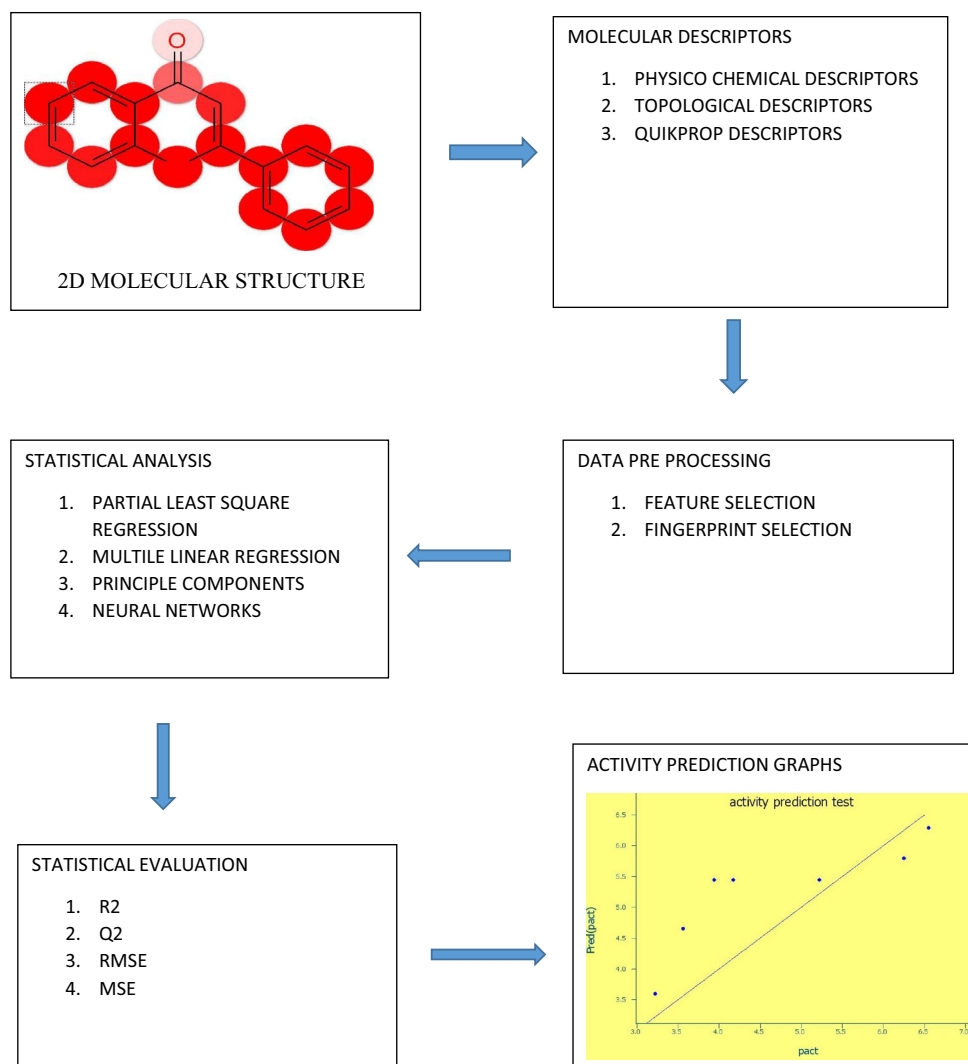
Introduction

Molecular similarity is one of the highly applied concepts in rational drug design. It assumes that structurally similar fragments will elicit a similar biological response. The concept of bioisosterism is closely related to molecular similarity where substructures may be interchanged with retention of some degree of biological activity. Assessment of molec-

ular similarity is a tedious process. Molecular graphs and molecular fingerprints can be used systematically to find special chemical features responsible for the biological activity. Fingerprints are binary vectors representing specific key structures in a chemical entity. Each bit score is represented by a binary score of either 1 or 0, where 1 represents the presence of a specific chemical feature, while 0 represents the absence of such feature. Computationally advanced software systems use either 32- or 64-bit systems. In 32-bit systems, the probability of encountering the same fragment is $1/2^{32}$ and for 64-bit applications the probability is $1/2^{64}$. The probability of collision is minimized by using

advanced computational software which gives statistically significant results. Molecular similarity may be transcribed into a numerical value and can be applied for similarity measurement, virtual screening methods and cluster analysis. In the present study, five different types of binary fingerprinting methods were used to predict the structural requirements for the selective inhibition of Tankyrase I on a series of substituted flavones. The suitability of the particular fingerprint for the selected molecules will be assessed by kernel-based partial least square values. Although Tankyrases have become attractive targets for anticancer agents, there are few effective drugs that inhibit Tankyrase. This is why this simulation study was carried out.

Cancer is one of the most concerning diseases of the modern world. All the cancers have a common characteristic feature known as a dysregulated cell cycle machinery. Among all cancers, the colorectal cancer (crc) has attracted our attention. The recent literature reports that colorectal cancer (crc) has the second highest mortality rate in the cancer segment [1]. The exact molecular mechanisms for the development of colorectal cancer have been deduced. The single major factor responsible for the development of colorectal cancer is the over activation of the Wnt signaling pathway by its central activator beta-catenin [2,3]. Under normal conditions, β-catenin levels are highly regulated by a feedback mechanism involving a β-catenin degradation complex [4]. Gene mutation causes the disruption of β-catenin degradation components. This leads to translocation of β-catenin into the nucleus, causing abnormal activation of transcription factors and gene networks responsible for the development of colorectal cancers. There is an urgent need to develop small molecule inhibitors for the selective inhibition of the Wnt signaling pathway, thereby increasing the levels of β-catenin degradation complex to combat colorectal cancer.

Tankyrases are members of the poly ADP-ribose polymerase (PARP) family of proteins. They have attracted attention because of their role in axin down regulation and stabilization of β-catenin. The post-translational modification involves the cleavage of PARP proteins, which results in splitting of NAD into ADP ribose and nicotinamide units. Poly ADP-ribose polymerases are responsible for many important biochemical signaling process in a cell machinery. The recently developed Tankyrase inhibitors were able to increase axin levels and down regulate β-catenin levels [5].

Recent literature also gave us insight about the advantage of selective inhibition of Tankyrase I. Partial knockdown of Tankyrase I leads to the shortening of telomere length [6]. The combination of selective Tankyrase I inhibitors with other class of drugs could be a viable strategy for the treatment of cancers like colorectal cancers.

Extensive literature search for small molecule inhibitors of Tankyrase I arouse our interest in the flavone pharmacophore due its diverse biological activity, and simple scaffold Flavones are naturally occurring secondary plant metabolites categorized under the broad class of flavonoids. They possess various degrees of free radical scavenging properties and are present in a wide variety of edible plants and vegetables. Flavonoids have also been shown to possess anti-tumor effect in various cancer cell lines. Inhibition of TNKS1 with flavone and its antiproliferative properties have already been reported. The present study is based on work reported in the literature [7].

Materials and methods

Selection of data set

In the present study, a data set of 25 out of 30 compounds was chosen from the literature [7] based on their structural diversity and activity. A training set of 19 molecules (70% of total molecules) was used to generate a kernel-based partial least square regression equation. The training set molecules were selected based on their structural diversity, activity range of 3 log order difference and activities covering the entire range.

To assess the predictive accuracy of the generated kernel-based regression model, a set of 6 molecules were chosen for the test set. The test set was selected in such a way that it is a representation of the training set. The training set was used to generate a 2D QSAR model, and the test set was used to validate the generated model.

Activity values (IC50) from the literature were converted to pIC50 (logarithmic scale) using options available in the calculator. The logarithmic activity is termed as *pact*.

Importing the energy minimized structures along with their logarithmic scale activity

The structures were imported into Maestro [8], screened for errors, valence parameters, invalid chemistry and duplicate structures in order to avoid interferences in the generation of the model.

Selection and incorporation of various molecular properties

Molecular descriptors were incorporated using the option molecular properties available with the Canvas 2.9 interface. There are main four types of descriptors in Canvas, namely physicochemical descriptors, topological descriptors, ligfilter descriptors and Qik-prop descriptors.

Feature selection

Feature selection is based on a hierarchical clustering to identify subsets of properties that are representative of a larger set. Sixty descriptors were incorporated and explained briefly.

1. *Adsorbability index (AI)* A molecular descriptor, which predicts the activated carbon adsorption of chemical substances from aqueous solutions. The adsorbability index [9] for a molecule is calculated by the expression:

$$AI = \sum_i A_i + \sum_i I_i$$

   where A represents the atomic or group contributions of increasing or decreasing adsorbability in the chemical species and I represents the necessary correction factors.

2. *ALOGP* One of the most applied universal lipophilicity descriptors derived directly from experimental data.

3. *Atomic composition indices* The descriptor possessing zero-dimensional attribute deduced directly from the composition of chemical entities and also provides information on the molecular weight and atomic composition of the chemical entity.

4. *Total information index on atomic composition* (I_{AC}) This descriptor calculates total information on atomic composition of a molecule directly from its molecular formula including hydrogen atoms [10].
   The atomic composition index (I_{AC}) is calculated using the formula:

$$I_{AC} = A^h \cdot \log_2 A^h - \sum_g A_g \cdot \log_2 A_g$$

   A^h total sum of atoms including hydrogens, A_g total sum of atoms which belongs to the gth chemical element.

5. *Nuclear information content descriptor* (I_{NUCL}) An important descriptor which gives the information on the total number of protons and neutrons present in all the types of nuclei in a given molecule.
   It is expressed by the following formula:

$$I_{NUCL} = \sum_{i=1}^{A} I_i^{n,p}$$

6. *Information index on size* (I_{SIZE}) This descriptor provides complete information content based on the atomic number [11].
   It is expressed as:

$$I_{SIZE} = A^h \cdot \log_2 A^h$$

A^h is the atom number with or without considering hydrogen atoms.

7. *Autocorrelation descriptors* (AC$_L$) This molecular descriptor is based on the autocorrelation function expressed as

$$AC_L = \int_a^b f(x) \cdot f(x+l) \cdot dx$$

8. *Topological electronic descriptors* (T^E) These descriptors [12] are based on partial atomic charges (q) and are expressed as

$$T^E = \sum_{i=1}^{A-1} \sum_{j=i+1}^{A} \frac{|q_i - q_j|}{r_{ij}^2}$$

9. *Partial charge weighted topological electronic index* (PCWT^E) A molecular descriptor based on the topological electronic index. It is expressed as

$$PCWT^E = \frac{1}{Q_{max}^-} \cdot \sum_{b=1}^{B} \left(\frac{|q_i - q_j|}{r_{ij}^2} \right)_b$$

10. *Local dipole index (D)* This descriptor calculates average differences in the magnitude of charge between overall bonded atom pairs (i–j) [13] and is expressed as

$$D = \frac{\sum_b |q_i - q_j|_b}{B}$$

11. *Atom in structure invariant index (ASII)* It belongs to a class of charge-related indices derived from hydrogen depleted molecular graphs. It is grouped under the category of global descriptors.
   It is expressed as

$$ASII_i = ASII_i^0 - h_i + q_i$$

12. *Charged partial surface area descriptors (CPSA)* These are the set of descriptors [14] that correlate shape and electronic information content to identify chemical compounds and to quantitate the polar interaction between molecules. They mainly consider the Vander Walls radius as a key feature for the quantitation.

13. *Partial negative surface area* (PNSA$_1$) It describes the information on total area of solvent-accessible surface area, comprised of all the negatively charged atoms.
   It is expressed as

$$PNSA_1 = \sum_{a-} SA_a^-$$

14. *Partial positive surface area* (PPSA$_1$) It is the total summation of the solvent-accessible surface area of all the positively charged atoms.
    It is expressed as

    $$PPSA_1 = \sum_{a+} SA_a^+$$

    The sum is restricted to positively charged atoms ($a+$)

15. *Total charge weighted negative surface area* (PNSA$_2$) The product of partial negative solvent-accessible surface area to the total negative charge (Q^-).
    It is expressed as

    $$PNSA_2 = Q^- \cdot \sum_{a-} SA_a^-$$

16. *Total charge weighted positive surface area* (PPSA$_2$) The product of partial positive solvent-accessible surface area multiplied by the total positive charge (Q^+).
    It is expressed as

    $$PPSA_2 = Q^+ \cdot \sum_{a+} SA_a^+$$

17. *Atomic charge weighted negative surface area* (PNSA$_3$) The product of atomic solvent-accessible surface areas and partial charges q_a^- over all negatively charged atoms.
    It is expressed as

    $$PNSA_3 = \sum_{a-} q_a^- \cdot SA_a^-$$

18. *Atomic charge weighted positive surface area* (PPSA$_3$) The summation of the products of atomic solvent-accessible surface areas and partial charges q_a^+ over all positively charged atoms.
    It is expressed as

    $$PPSA_3 = \sum_{a+} q_a^+ \cdot SA_a^+$$

19. *Difference in the charged partial surface area* (DPSA$_1$) The difference between partial positive solvent-accessible surface area and partial negative solvent-accessible surface area.
    It is expressed as

    $$DPSA_1 = PPSA_1 - PNSA_1$$

20. *Difference in the total charge weighted surface area* (DPSA$_2$) The difference between total charge weighted positive solvent-accessible surface area and total charge weighted negative solvent-accessible surface area.

It is expressed as

$$DPSA_2 = PPSA_2 - PNSA_2$$

21. *Difference in the atomic charge weighted surface area* (DPSA$_3$) The difference between atomic charge weighted positive solvent-accessible surface area and atomic charge weighted negative solvent-accessible surface area.
    It is expressed as

    $$DPSA_3 = PPSA_3 - PNSA_3$$

22. *Relative negative charge (RNCG)* Partial charge of the most negative atom divided by total negative charge.
    It is expressed as

    $$RNCG = \frac{Q_{max}^-}{Q^-}$$

23. *Relative positive charge (RPCG)* Partial charge of the most positive atom divided by total positive charge.
    It is expressed as

    $$RPCG = \frac{Q_{max}^+}{Q^+}$$

24. *Relative negatively charged surface area (RNCS)* The solvent-accessible surface area of the most negative atom divided by the relative negative charge (RNCG).
    It is expressed as

    $$RNCS = \frac{SA_{max}^-}{RNCG}$$

25. *Relative positively charged surface area (RPCS)* The solvent-accessible surface area of the most positive atom divided by the relative positive charge (RPCG).
    It is expressed as

    $$RPCS = \frac{SA_{max}^+}{RPCG}$$

26. *Total hydrophobic surface area (TASA)* The sum of solvent-accessible surface areas of atoms with absolute value of partial charges less than 0.2.
    It is expressed as

    $$TASA = \sum_a SA_a$$

27. *Total polar surface area (TPSA)* The sum of solvent-accessible surface areas of atoms with absolute value of partial charges greater than or equal to 0.2.

It is expressed as

$$TPSA = \sum_a SA_a$$

28. *Relative hydrophobic surface area (RASA)* The ratio of total hydrophobic surface area (TASA) to the total molecular solvent-accessible surface area (SASA). It is expressed as

$$RASA = \frac{TASA}{SASA}$$

29. *Relative polar surface area (RPSA)* The total polar surface area (TPSA) divided by the total molecular solvent-accessible surface area (SASA). It is expressed as

$$RPSA = \frac{TPSA}{SASA}$$

30. *RHTA index* The ratio of hydrogen bond donor groups to hydrogen bond acceptor groups. It is expressed as

$$RHTA = \frac{HBD}{HBA}$$

31. *SSAH index* The total surface area of hydrogen atoms that can be readily donated. It is expressed as

$$SSAH \equiv HDSA = \sum_d SA_d$$

32. *RSAH index* The average surface area of hydrogen atoms that can be donated. It is expressed as

$$RSAH = \frac{\sum_d SA_d}{HBD}$$

33. *RSHM index* The fraction of the total molecular surface area associated with hydrogen atoms that can be readily donated. It is expressed as

$$RSHM \equiv FHDSA = \frac{\sum_d SA_d}{SASA}$$

34. *SSAA index* The sum of the surface areas of all hydrogen bond acceptor atoms. It is expressed as

$$SSAA \equiv HASA = \sum_a SA_a$$

35. *RSAM index* The fraction of the total molecular surface area associated with H-bond acceptor groups. It is expressed as

$$RSAM \equiv FHASA = \frac{\sum_a SA_a}{SASA}$$

36. *HDCA index* The sum of charged surface areas of hydrogen atoms that can be donated. It is expressed as

$$HDCA = \sum_d q_d \cdot SA_d$$

37. *HBSA index* The sum of the surface areas of both hydrogen atoms that can be donated to hydrogen acceptor atoms. It is expressed as

$$HBSA = HDSA + HASA$$

38. *Graph distance complexity* Molecular descriptor derived from the distance matrix D [15]. It is expressed as

$$H_D = \sum_{i=1}^{A} \frac{\sigma_i}{I_{ROUV}} \cdot v_i^d = \sum_{i=1}^{A} \frac{\sigma_i}{2W} \cdot v_i^d$$

39. *Polar hydrogen factor* (Q_H) This descriptor correlates polarity of molecules to C–H bonds. Its application is limited to halogenated hydrocarbons. It is expressed as

$$Q_H = \sum_b \left[\sum_C k_C + \sum_\alpha k_\alpha + \sum_\beta k_\beta \right]$$

40. *Q polarity index* Topological polarity index derived from the electro topological intrinsic state of the atoms confined in a molecule. It is expressed as

$$Q = \frac{A^2 \cdot \sum_{i=1}^{A} I_i^{ALK}}{\left(\sum_{i=1}^{A} I_i \right)^2}$$

41. *Molecular polarizability effect index (MPEI)* This descriptor works by the principle that molecules are polarized by electrostatic potential fields [16]. The index is calculated by summing the polarizability contributions from different atoms in a molecule.

It is expressed as

$$MPEI = \sum_{i=1}^{A} PEI_i$$

42. *Balaban distance connectivity index (J)* It is one of the topological descriptors and its value does not vary substantially with the size of the molecules or number of ring systems.
It is expressed as

$$J = \frac{B}{C+1} \cdot \sum_b (\sigma_i \cdot \sigma_j)_b^{-1/2}$$
$$= \frac{1}{C+1} \cdot \sum_b (\bar{\sigma}_i \cdot \bar{\sigma}_j)_b^{-1/2}$$

43. *Atomic charge (q)* The experimental approach to calculate atomic charge is called Mulliken population analysis [17]. The method allocates electrons to atoms, transforming atomic charge to a local descriptor.
It is expressed as

$$q_a = Z_a - \sum_{\mu=1}^{N_{AO}} \sum_{v=1}^{N_{AO}} P_{\mu v} \cdot S_{\mu v}$$

44. *Sub-molecular polarity parameter* (SPP$^1\Delta$): An electronic descriptor [18] that mathematically determines the excess charge difference between a pair of atoms.
It is expressed as

$$^1\Delta = \left| Q_{max}^+ - Q_{max}^- \right|$$

45. *Second-order sub-molecular polarity parameter* ($^2\Delta$): Determines the second prime difference of excess charges [19].
It is expressed as

$$DP = \frac{\left| Q_{max}^+ - Q_{max}^- \right|}{r_{\pm}^2} = \frac{^1\Delta}{r_{\pm}^2}$$

46. *Molar polarization* (P_M) The dipole moment induced for each unit volume V is termed molar polarization. Clausius–Mossotti equation explains this descriptor.
It is expressed as

$$P_M = \frac{\varepsilon - 1}{\varepsilon + 2} \cdot \frac{MW}{\varrho} = \frac{4\pi}{3} \cdot N_A \cdot \alpha \cdot$$
$$E = \frac{n_D^2 - 1}{n_D^2 + 2} \cdot \frac{MW}{\varrho} = MR$$

47. *Atom–atom polarizability* A chemical reactivity index solely calculated on the basis of perturbation theory.
It is expressed as

$$P_{ab} = 4. \sum_i \sum_j \sum_\mu \sum_v \frac{C_{i\mu,a} \cdot C_{j\mu,a} \cdot C_{iv,b} \cdot C_{jv,b}}{\varepsilon_i - \varepsilon_j}$$

48. *Anisotropy of the polarizability* It quantifies the deviation of molecular polarizability from an equivalent spherical shape.
It is expressed as

$$\beta^2 = \frac{(\alpha_{xx} - \alpha_{yy})^2 + (\alpha_{yy} - \alpha_{zz})^2 + (\alpha_{zz} - \alpha_{xx})^2}{2}$$

49. *Overall electronic constants* These are the Hammett substitution constants [20] which measure the total electronic effect of meta and para substituted benzene rings comprised of substituents in a side chain.
It is expressed as

$$\sigma_{m,p} = \frac{1}{\varrho} \cdot \log\left(\frac{K^X}{K_0}\right) = \frac{1}{\varrho} \cdot \left(pK_a^0 - pK_a^X\right)$$

50. *Information index on the molecular symmetry* (I_{SYM}): An important molecular symmetry descriptor which is solely based on total information content.
It is expressed as

$$I_{SYM} = A \cdot \log_2 A - \sum_{g=1}^{G} A_g \log_2 A_g$$

51. *Joshi steric descriptor* (JM$_1$) A descriptor that directly measures the steric effect [21] of substituents. It is grouped under the broad class of quantum chemical descriptors.
It is expressed as

$$JM1 = \frac{\Delta E_X}{\Delta E_H} \quad \log(JM1) = \log(\Delta E_X) - \log(\Delta E_H)$$

52. *Substituent front strain* (S_f) A steric descriptor [22] obtained using empirical force fields and calculated based on enthalpy of formation.
It is expressed as

$$S_f = \Delta H_f^0 [XC(CH_3)_3] - \Delta H_f^0 [XCH_3]$$
$$+ 8.87 \, [10^4 \, J/mol]$$

53. *Steric vertex topological descriptor (SVTI)* Best performing steric descriptor for alkyl groups [23] char-

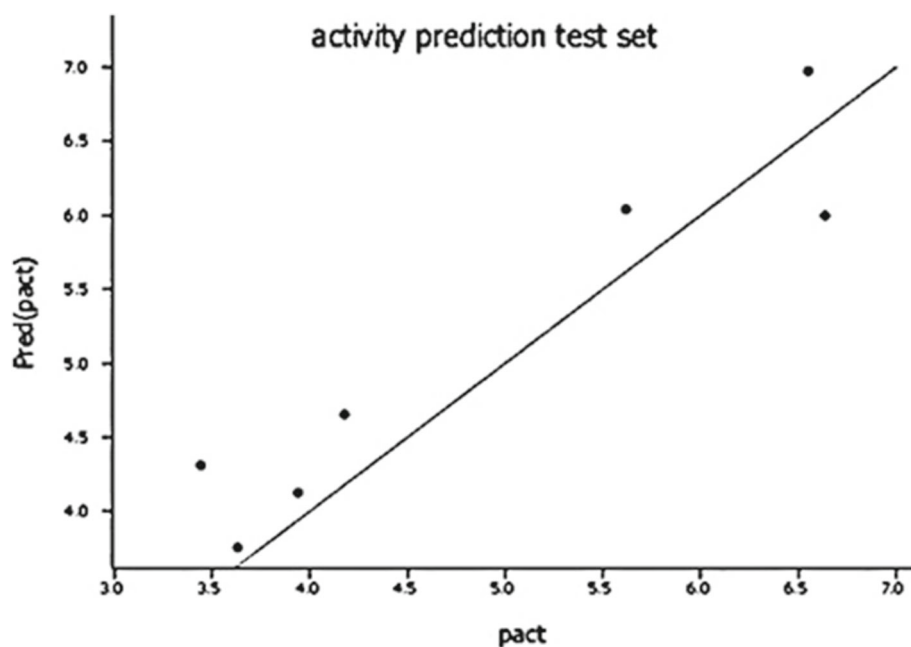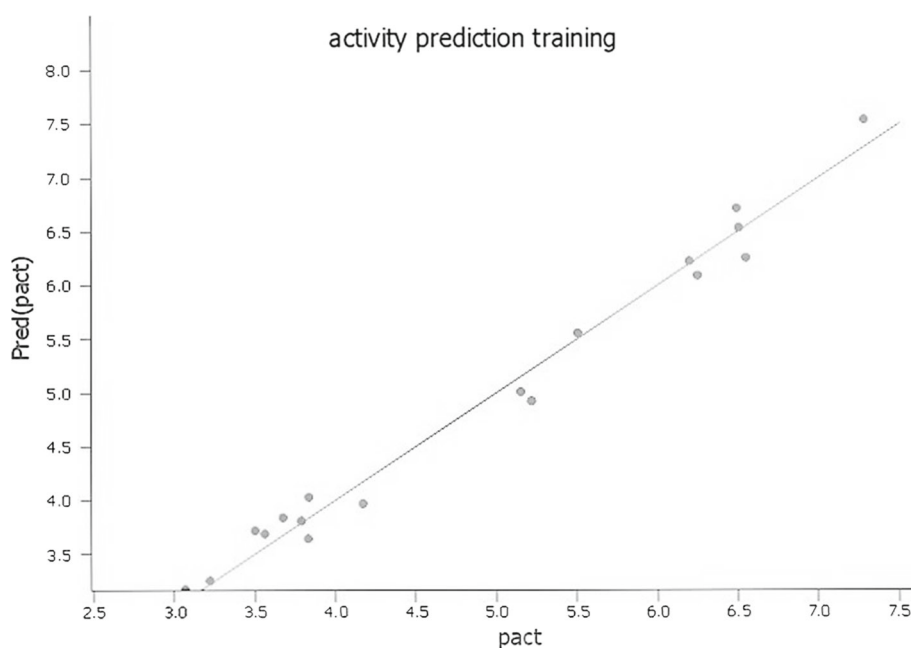**Fig. 1** Actual versus predicted test set (fingerprint—atom pairs)

Fig. 2 Actual versus predicted training set (fingerprint—atom pairs)

acterized by their topological distance (d) from an H-depleted molecular graph.

It is expressed as

$$\text{SVTI} = \sum_{j=1}^{A_X} d_{ij} \quad \forall d_{ij} \leq 3$$

54. *Steric density descriptor* (SD_X) The substituent steric descriptor [24], which correlates molecular mass to van der Waals volume.

It is expressed as

$$\text{SD}_X = \left(\frac{MW}{V_{\text{VDW}}}\right)_X - \left(\frac{MW}{V_{\text{VDW}}}\right)_H$$
$$= \left(\frac{MW}{V_{\text{VDW}}}\right)_X - 0.29$$

55. *Model of the Frontier steric effect descriptor* (R_S) Theoretical descriptor used to estimate Taft's steric constant [25] which is based on the fundamental characteristics of constituent atoms.

Table 1 KPLS results on training set

KPLS factors	SD	R^2
1	0.8362	0.682
2	0.4173	0.9261
3	0.2066	0.9832

Table 2 KPLS results on test set

KPLS factors	RMS deviation	Q^2
1	0.6871	0.7107
2	0.4920	0.8516
3	0.5035	0.8446

Table 3 Activity prediction (actual vs. predicted)

Structure	Model set	Pact	Pred (pact)	Error
tan1	Training	6.495	6.708	0.214
tan2	Training	5.509	5.548	0.04
tan3	Test	5.62	6.037	0.417
tan4	Test	6.553	6.972	0.419
tan5	Training	6.252	6.081	−0.171
tan6	Training	6.201	6.219	0.018
tan7	Training	6.509	6.526	0.017
tan8	Test	6.638	5.994	−0.645
tan9	Training	6.553	6.25	−0.303
tan10	Training	7.284	7.534	0.25
tan11	Training	3.678	3.833	0.156
tan12	Training	3.225	3.249	0.024
tan13	Test	3.444	4.306	0.863
tan14	Test	3.633	3.75	0.118
tan15	Training	3.504	3.712	0.207
tan16	Training	3.071	3.167	0.096
tan17	Test	4.18	4.649	0.468
tan18	Training	5.222	4.919	−0.303
tan19	Training	4.174	3.962	−0.212
tan20	Training	3.565	3.685	0.119
tan21	Training	3.79	3.805	0.014
tan22	Training	3.836	3.637	−0.199
tan23	Training	3.839	4.022	0.184
tan24	Training	5.155	5.005	−0.15
tan25	Test	3.943	4.12	0.177

It is expressed as

$$R_S = -30 \cdot \log \left(1 - \sum_{i=1}^{n} \frac{R_i^2}{4 \cdot r_i^2} \right)$$

56. *Carbo similarity index (C)* This descriptor [26] compares two molecules based on their electron density.

It is also applied to compare any structural properties between molecules.
It is expressed as

$$C_{st} = \frac{\sum_{k=1}^{N} P_{sk} \cdot P_{tk}}{\left(\sum_{k=1}^{N} P_{sk}^2 \right)^{1/2} \cdot \left(\sum_{k=1}^{N} P_{tk}^2 \right)^{1/2}}$$

57. *Electrophilic atomic frontier electron density descriptor* (f_a^-) Molecular descriptor reflecting the electron density status in the HOMO orbital of a compound.
It is expressed as

$$f_a^- = \sum_{\mu} \left(c_{\mathrm{HOMO}, \mu} \right)^2$$

58. *Hardness indices* (η) These are the class of descriptors derived directly from density functional theory [27]. They correspond to the second-order derivative energy levels with respect to the total number of electrons in the system.
It is expressed as

$$\eta = \frac{1}{2} \left(\frac{\partial^2 E}{\partial N_{\mathrm{el}}^2} \right)_{v(r)} = \left(\frac{\partial \mu}{\partial N_{\mathrm{el}}} \right)_{v(r)}$$
$$= \int h(r) \mathrm{d}r = \frac{1}{2 \cdot S}$$

59. *Composite nuclear potential* $(v(r))$ This descriptor defines the pattern of the nuclei of a molecule.
It is expressed as

$$v(r) = \sum_{a=1}^{A} \frac{Z_a}{|r - R_a|}$$

60. *Kier alpha molecular flexibility index* (Φ) Direct measurement of molecular flexibility [28] derived from Kier alpha adapted shape descriptors.
It is expressed as

$$\Phi = \frac{{}^1 K_\alpha \cdot {}^2 K_\alpha}{A}$$

Incorporation of Binary fingerprints

The seven available binary fingerprints were incorporated using Canvas interface. Only the models with significant contribution were retained for further process.

Kernel-based partial least square regression using different binary fingerprints

1. Binary fingerprint—*atom pairs* [29]

Description: represents pair of atoms, which are differentiated into atom types and distance of separation.

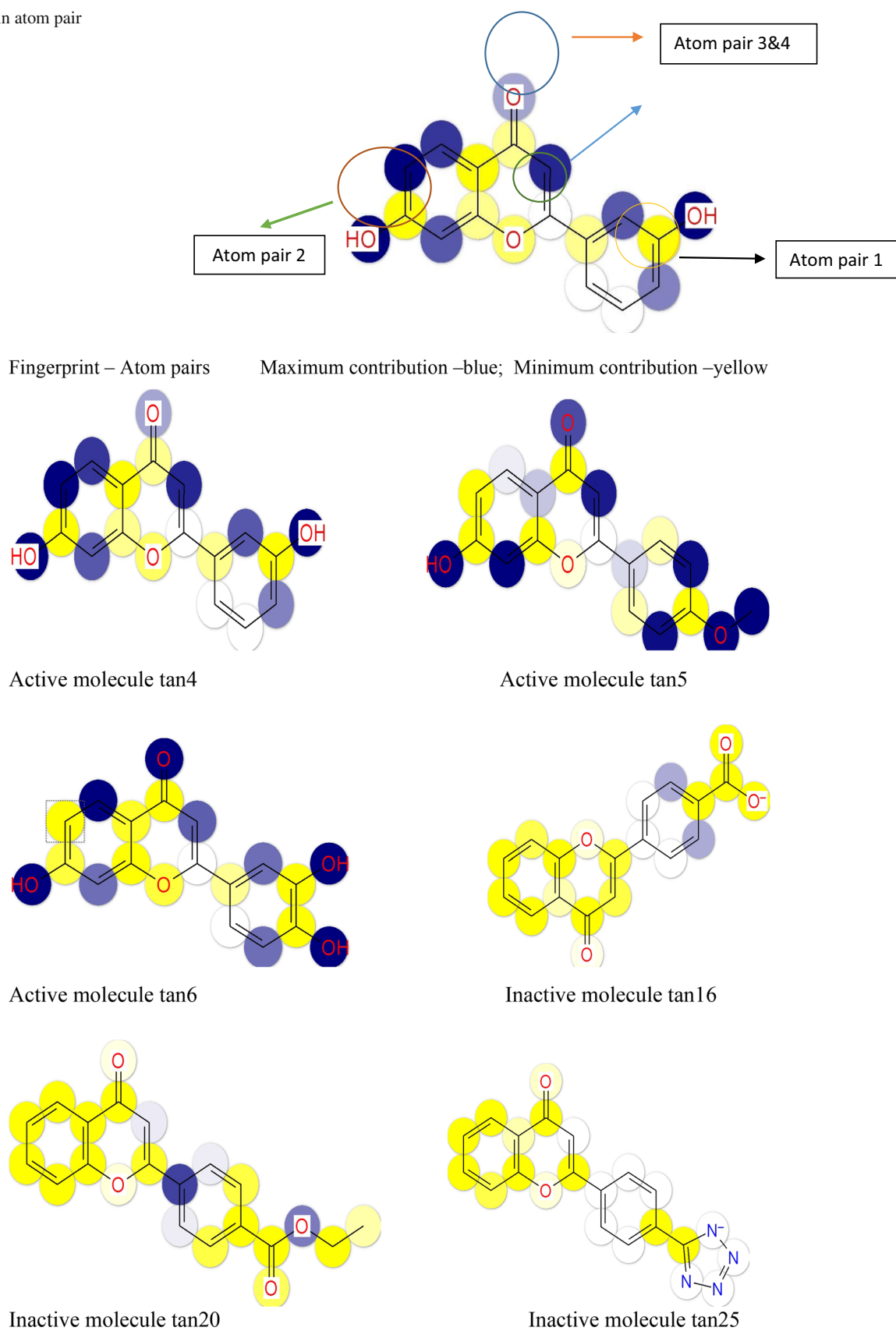**Fig. 3** Hashing in atom pair fingerprinting

Fingerprint – Atom pairs   Maximum contribution –blue; Minimum contribution –yellow

Active molecule tan4

Active molecule tan5

Active molecule tan6

Inactive molecule tan16

Inactive molecule tan20

Inactive molecule tan25

Fig. 4 Atomic contribution model for fingerprint—atom pairs

Fig. 5 Actual versus predicted training set (fingerprint –atom triplets)

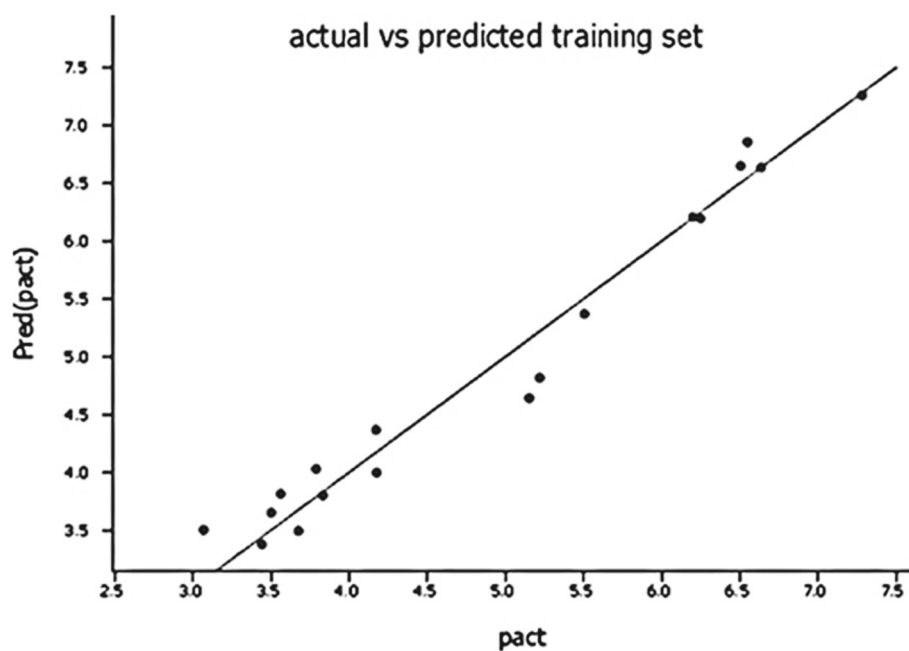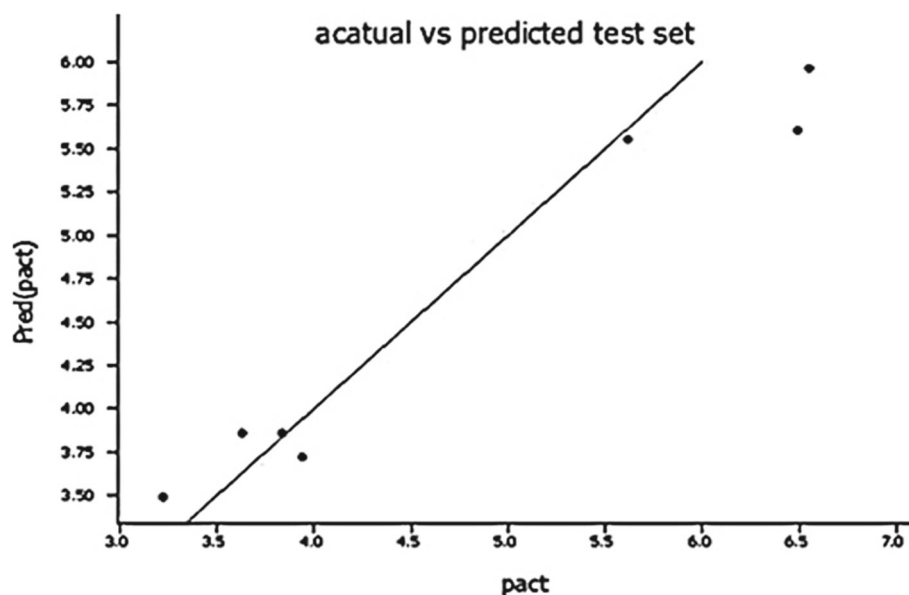**Fig. 6** Actual versus predicted test set (fingerprint—atom triplet)

Table 4 KPLS results on training set

KPLS factors	SD	R^2
1	0.7597	0.731
2	0.4084	0.9274
3	0.2768	0.9690

Table 5 Results of KPLS on test set

KPLS factors	RMS deviation	Q^2
1	0.9022	0.5293
2	0.5131	0.8478
3	0.4246	0.8908

The concept of atom pairs has been used in this fingerprint. A molecular entity is hashed into smaller fragments to give an integer value, and the shortest distance between any two atoms is given by the term d. The actual contribution of the model to the activity is assessed by kernel-based partial least square equation by considering their regression values (R^2 and Q^2 values).

2. Binary fingerprint—*atom triplets* [29]

Description: triplets of atoms and three distances separating them.

The model is an extension of the atom-pair fingerprint. Triplets represent three atoms and the distance at which they are separated. A triplet can be presented in 6 different ways according to the theory of permutation and combination. An operation is performed to exclude the bits which correspond only to the permutation Typea-dab-Typeb-dbc-Typec-dca. The assessment method is identical to the description under atom pairs.

3. Binary fingerprint—*fp linear* [29]

Description: linear fragments + ring closures.

The linear fingerprint fragments a molecule in every possible linear combination path. The default application considers up to 7 bonds. A hashing operation is performed for each linear fragment to generate a corresponding bit address. To apply a linear path on closed ring systems, linear paths may be extended up to 14 bonds. The linear fingerprints can be extensively applied to compounds with extended ring systems. The assessment method used is the same for atom pairs.

4. Binary fingerprint—*2D molprint* [29]

Description: The molprint fingerprint uses heavy atoms present in a molecule and surrounding environment which is separated by a maximum of two bonds. The bit value is measured from a stored data containing a heavy atom and distance separated by other atoms by one or two bond orders. The assessment method used is the same for atom pairs.

5. Binary fingerprint—*fp dendritic* [29]

Description: Linear and branched fragments.

The dendritic fingerprint uses a combination of both linear and branched fragments up to a user-defined value of 5 bonds. There is no special treatment for ring systems which are considered as branched fragments. The assessment method used is the same for atom pairs.

Results and discussion

1. Results of kernel-based partial least square [30] regression on fingerprint—atom pairs (Figs. 1, 2)

The fingerprint atom pairs gave a statistically significant 2D QSAR model with excellent regression coefficient values and cross-validation coefficient values as represented in Tables 1 and 2. The model also showed good predictive accuracy in both test and training set molecules as demonstrated in Table 3. A hashing technique used is represented in Fig. 3. For the assessment of atomic contribution to the model, three molecules were taken from each active and inactive set. Atoms positively contributing to activity were colored blue and atoms detrimentally contributing to the model were colored yellow which is shown in Fig. 4.

Table 6 Activity prediction (actual vs. predicted)

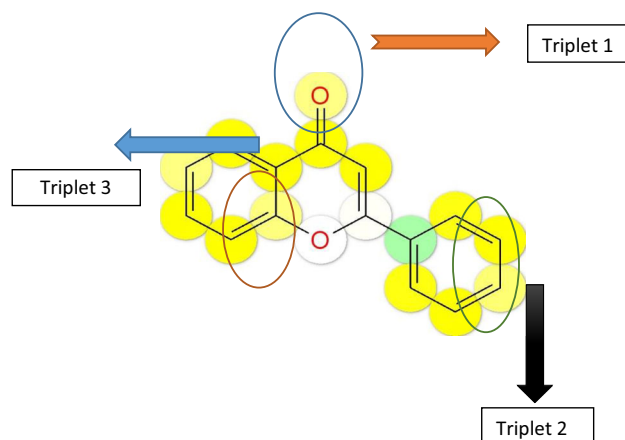| Structure	Model set	Pact	Pred (pact)	Error
tan1	Training	6.495	6.708	0.214
tan2	Training	5.509	5.548	0.04
tan3	Test	5.62	6.037	0.417
tan4	Test	6.553	6.972	0.419
tan5	Training	6.252	6.081	− 0.171
tan6	Training	6.201	6.219	0.018
tan7	Training	6.509	6.526	0.017
tan8	Test	6.638	5.994	− 0.645
tan9	Training	6.553	6.25	− 0.303
tan10	Training	7.284	7.534	0.25
tan11	Training	3.678	3.833	0.156
tan12	Training	3.225	3.249	0.024
tan13	Test	3.444	4.306	0.863
tan14	Test	3.633	3.75	0.118
tan15	Training	3.504	3.712	0.207
tan16	Training	3.071	3.167	0.096
tan17	Test	4.18	4.649	0.468
tan18	Training	5.222	4.919	− 0.303
tan19	Training	4.174	3.962	− 0.212
tan20	Training	3.565	3.685	0.119
tan21	Training	3.79	3.805	0.014
tan22	Training	3.836	3.637	− 0.199
tan23	Training	3.839	4.022	0.184
tan24	Training	5.155	5.005	− 0.15
tan25	Test	3.943	4.12	0.177

Fig. 7 Hashing in atom triplets

2. Results of kernel-based partial least square regression on fingerprint—atom triplets (Figs. 5, 6)

The fingerprint atom triplets gave a statistically significant 2D QSAR model with excellent regression coefficient val-
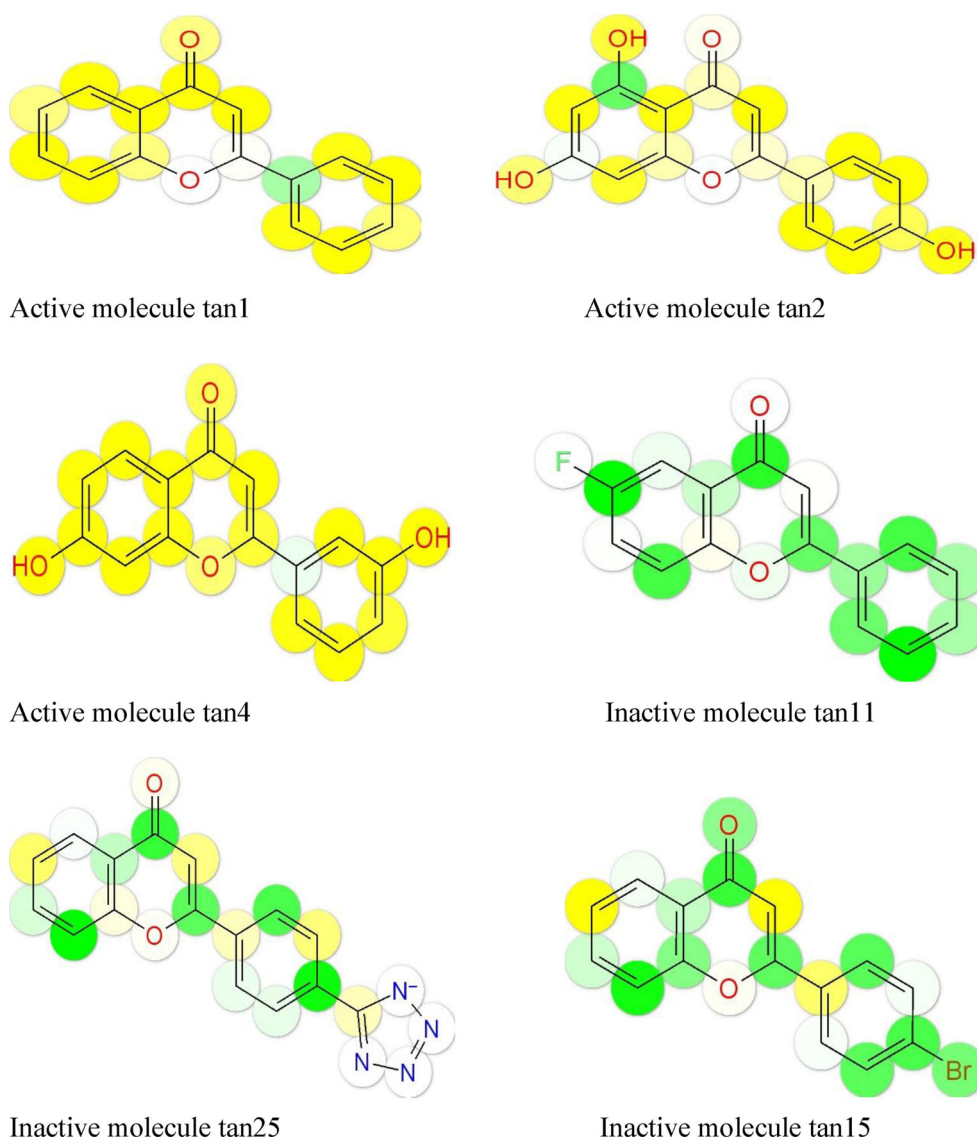
Active molecule tan1

Active molecule tan2

Active molecule tan4

Inactive molecule tan11

Inactive molecule tan25

Inactive molecule tan15

Fig. 8 Atomic contribution model for atom fingerprint—atom triplets

Table 7 KPLS results on training set

KPLS factors	SD	R^2
1	1.081	0.4748
2	0.3914	0.9358
3	0.1573	0.9904

Table 8 KPLS results on training set

KPLS factors	RMS deviation	Q^2
1	0.7408	0.6448
2	0.7716	0.6146
3	0.8714	0.5085

ues and cross-validation coefficient values as represented in Tables 4 and 5. The model also showed good predictive accuracy in both test and training set molecules as evidenced in Table 6. The hashing pattern in the atom triplet is shown in Fig. 7.
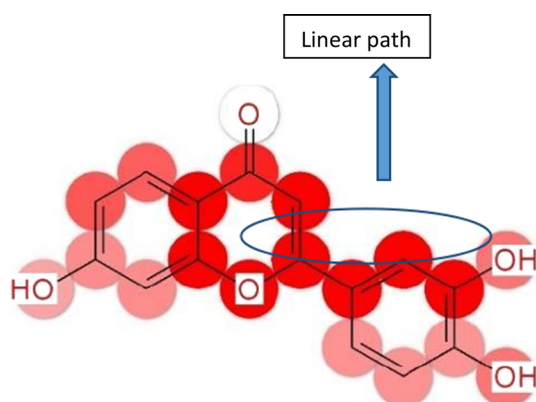
For the assessment of atomic contribution to the model, three molecules were taken from each active and inactive set,

atoms positively contributing to the activity were colored yellow and atoms detrimentally contributing to the model were colored green which is shown in Fig. 8.

The fingerprint fp linear gave a statistically significant 2D QSAR model with excellent regression coefficient values and cross-validation coefficient values as represented in Tables 7 and 8. The model also showed good predictive

Table 9 Activity prediction (actual vs. predicted)

Structure	Model set	Pact	Pred (pact)	Error
tan1	Training	6.495	6.401	− 0.093
tan2	Test	5.509	4.637	− 0.871
tan3	Training	5.62	5.484	− 0.136
tan4	Training	6.553	6.518	− 0.035
tan5	Test	6.252	5.454	− 0.798
tan6	Training	6.201	6.311	0.11
tan7	Test	6.509	6.862	0.353
tan8	Training	6.638	6.829	0.191
tan9	Training	6.553	6.428	− 0.124
tan10	Training	7.284	7.178	− 0.106
tan11	Training	3.678	3.509	− 0.168
tan12	Test	3.225	2.468	− 0.758
tan13	Training	3.444	3.51	0.067
tan14	Training	3.633	3.686	0.053
tan15	Training	3.504	3.572	0.068
tan16	Training	3.071	2.94	− 0.131
tan17	Training	4.18	3.996	− 0.184
tan18	Training	5.222	5.352	0.13
tan19	Test	4.174	5.241	1.067
tan20	Test	3.565	4.67	1.104
tan21	Training	3.79	4.099	0.309
tan22	Test	3.836	4.764	0.928
tan23	Training	3.839	3.943	0.104
tan24	Training	5.155	5.162	0.007
tan25	Training	3.943	3.884	− 0.06

Fig. 9 Hashing in fp linear fingerprint

ability in both test and training set molecules as evidenced in Table 9. Hashing pattern in fp linear is represented in Fig. 9.

3. Results of kernel-based partial least square regression on fingerprint—fp linear (Figs. 10, 11)

For the assessment of atomic contribution to the model [31], three molecules were taken from each active and inactive set, and atoms positively contributing to activity were colored red and atoms detrimentally contributing to the model were colored blue which is shown in Fig. 12.

4. Results of kernel-based partial least square regression on fingerprint—2D molprint (Figs. 13, 14)

The finger print 2D molprint gave a statistically insignificant 2D QSAR model with a large difference in regression coefficient values and cross-validation coefficient values as presented in Tables 10 and 11. 2D molprint is not a suitable fingerprint for the selected flavone class. The model also showed poor predictive accuracy in both test and training set molecules as shown in Table 12. Hashing pattern is represented in Fig. 15.

For the assessment of atomic contribution to the model, three molecules were taken from each active and inactive sets and atoms positively contributing to the activity were colored maroon and atoms detrimentally contributing to the model were colored blue as shown in Fig. 16.

5. Results of kernel-based partial least square regression on fingerprint—fp dendritic

The fingerprint fp dendritic gave a statistically insignificant 2D QSAR model with a large difference in the regression coefficient values and cross-validation coefficient values as represented in Tables 13 and 14. Therefore, fp dendritic [10] is not a suitable fingerprint for the selected flavone class (Figs. 17, 18, 19). The model also showed poor predictive ability in both test and training set molecules as evidenced in Table 15. The hashing pattern is represented in Fig. 20.

For the assessment of atomic contribution to the model, three molecules were taken from each active and inactive set, and atoms positively contributing to the activity were colored Orange and atoms detrimentally contributing to the model were colored maroon which is shown in Fig. 19.

Structure–activity relationship studies

Figure 21 shows that substitution with electron-donating groups (e.g., hydroxyl) on both rings A and C was found to increase activity in all the 5 fingerprint models. The only

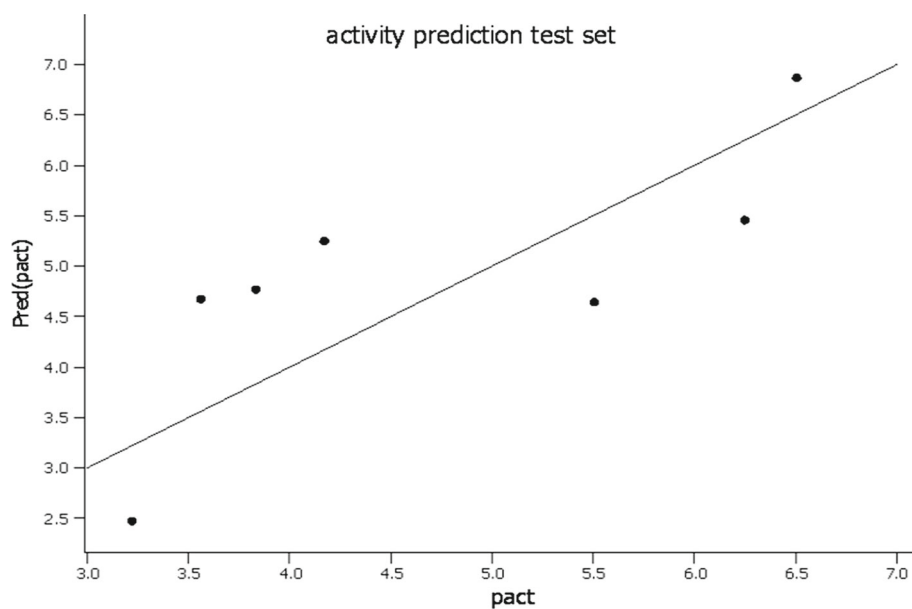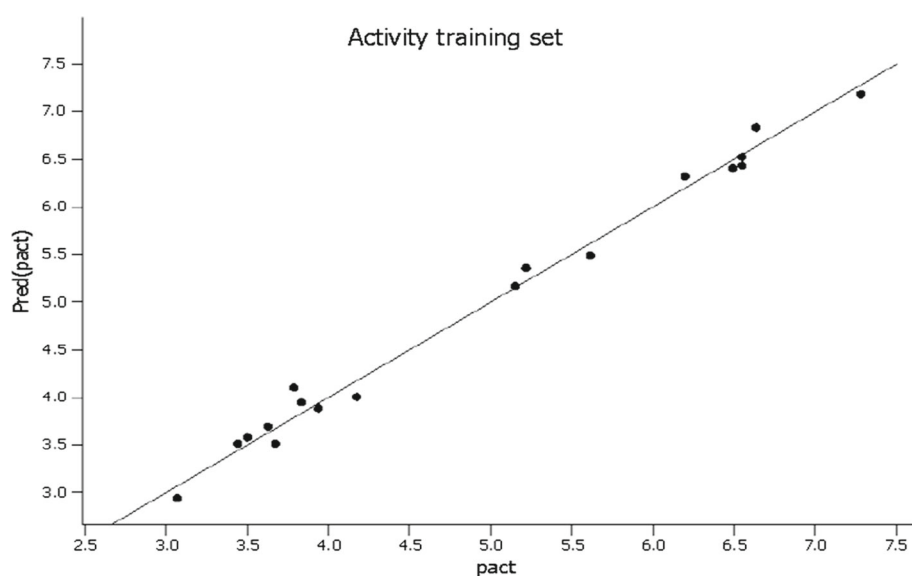**Fig. 10** Actual versus predicted test set (fingerprint—fp linear)

Fig. 11 Actual versus predicted training set (fingerprint—fp linear)

exception was molecule Tan 1, which is an active molecule even though it has no substitutions on both rings A and C. Substitution with electron withdrawing groups on both rings A and C was found to decrease activity in all the 5 fingerprint models. This may be attributed to the electronic parameters exerted by various functional groups on the aromatic rings.
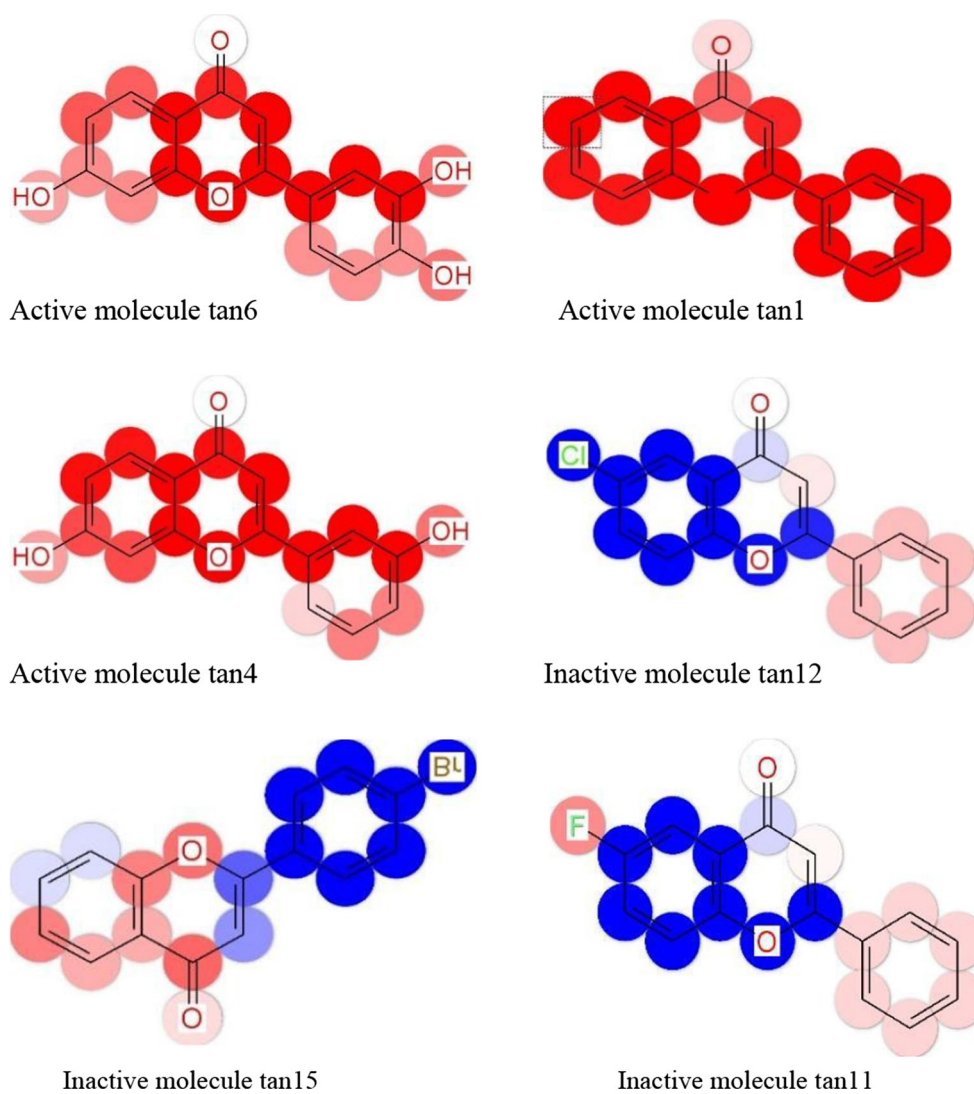
Active molecule tan6

Active molecule tan1

Active molecule tan4

Inactive molecule tan12

Inactive molecule tan15

Inactive molecule tan11

Fig. 12 Atomic contribution model for fingerprint—fp linear

Fig. 13 Actual versus predicted training set (fingerprint—2D molprint)

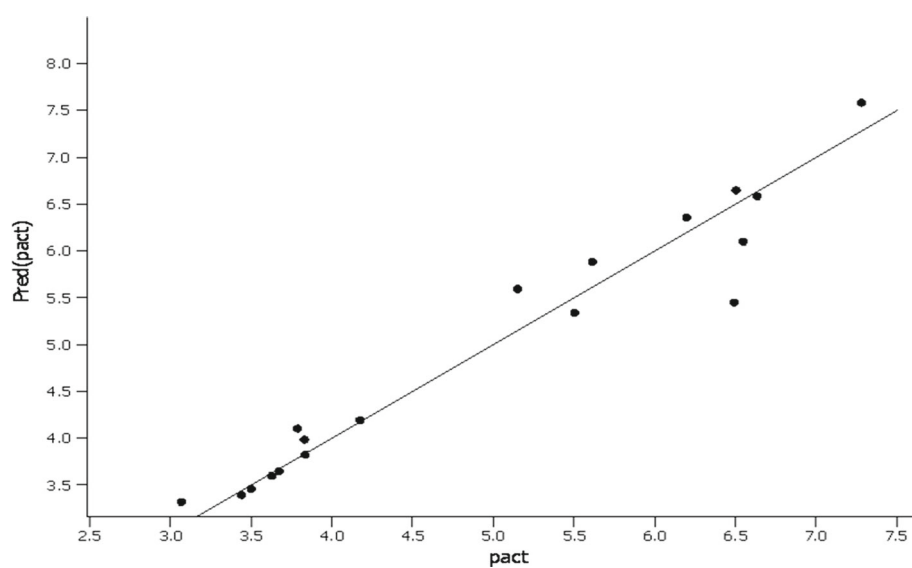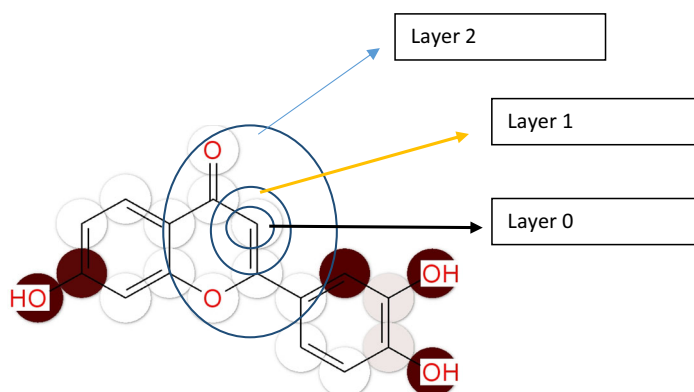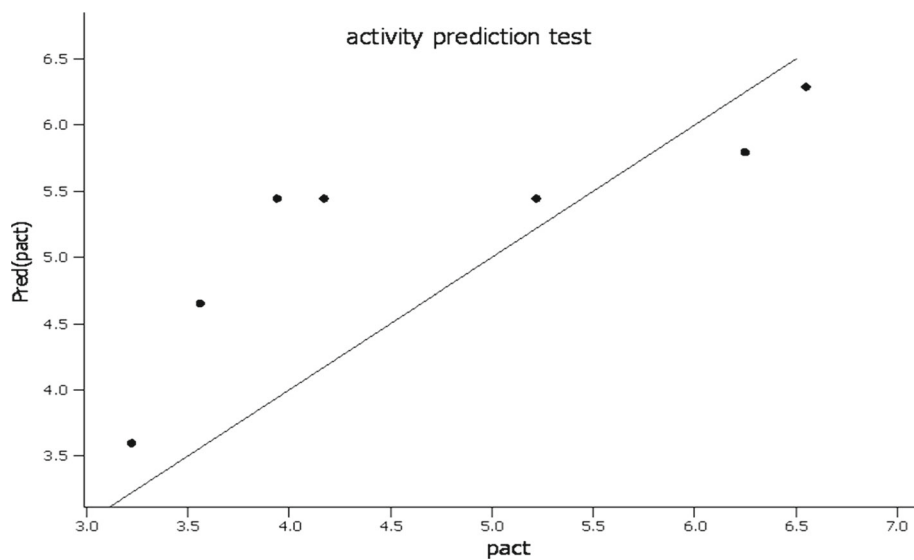**Fig. 14** Actual versus predicted test set (fingerprint—2D molprint)

LAYER	0	1	2
	C .aromatic(sp2)	C .aromatic (sp2)	C .aromatic(sp2)
		C. aromatic (sp2)	C .aromatic (sp2)
			C=O (sp2)
			C-O

Fig. 15 Hashing in 2D molprint (circular fingerprints)

Table 10 KPLS results on training set

KPLS factors	SD	R^2
1	1.038	0.5204
2	0.5195	0.8879
3	0.3841	0.9431

Table 11 KPLS results on test set

KPLS factors	RMS deviation	Q^2
1	0.6632	0.7047
2	0.7947	0.5706
3	0.8860	0.4278

Table 12 Activity prediction (actual vs. predicted)

Structure	Model set	Pact	Pred (pact)	Error
tan1	Training	6.495	5.442	− 1.053
tan2	Training	5.509	5.333	− 0.175
tan3	Training	5.62	5.876	0.257
tan4	Test	6.553	6.284	− 0.269
tan5	Test	6.252	5.791	− 0.46
tan6	Training	6.201	6.355	0.155
tan7	Training	6.509	6.641	0.132
tan8	Training	6.638	6.576	− 0.062
tan9	Training	6.553	6.09	− 0.463
tan10	Training	7.284	7.576	0.292
tan11	Training	3.678	3.641	− 0.037
tan12	Test	3.225	3.592	0.366
tan13	Training	3.444	3.386	− 0.058
tan14	Training	3.633	3.592	− 0.041
tan15	Training	3.504	3.452	− 0.052
tan16	Training	3.071	3.311	0.241
tan17	Training	4.18	4.19	0.009
tan18	Test	5.222	5.442	0.22
tan19	Test	4.174	5.442	1.268
tan20	Test	3.565	4.65	1.084
tan21	Training	3.79	4.097	0.306
tan22	Training	3.836	3.977	0.142
tan23	Training	3.839	3.817	− 0.022
tan24	Training	5.155	5.586	0.431
tan25	Test	3.943	5.442	1.499

During the analysis of binding site of flavones in the active site of Tankyrase 1, it is evident that hydrogen bond donor interactions are prominent for the effective binding. Conse-quently, electron-donating groups such as –OH and –NH2 were found to increase binding affinity toward Tankyrase I according to the developed QSAR model.

Active molecule tan2
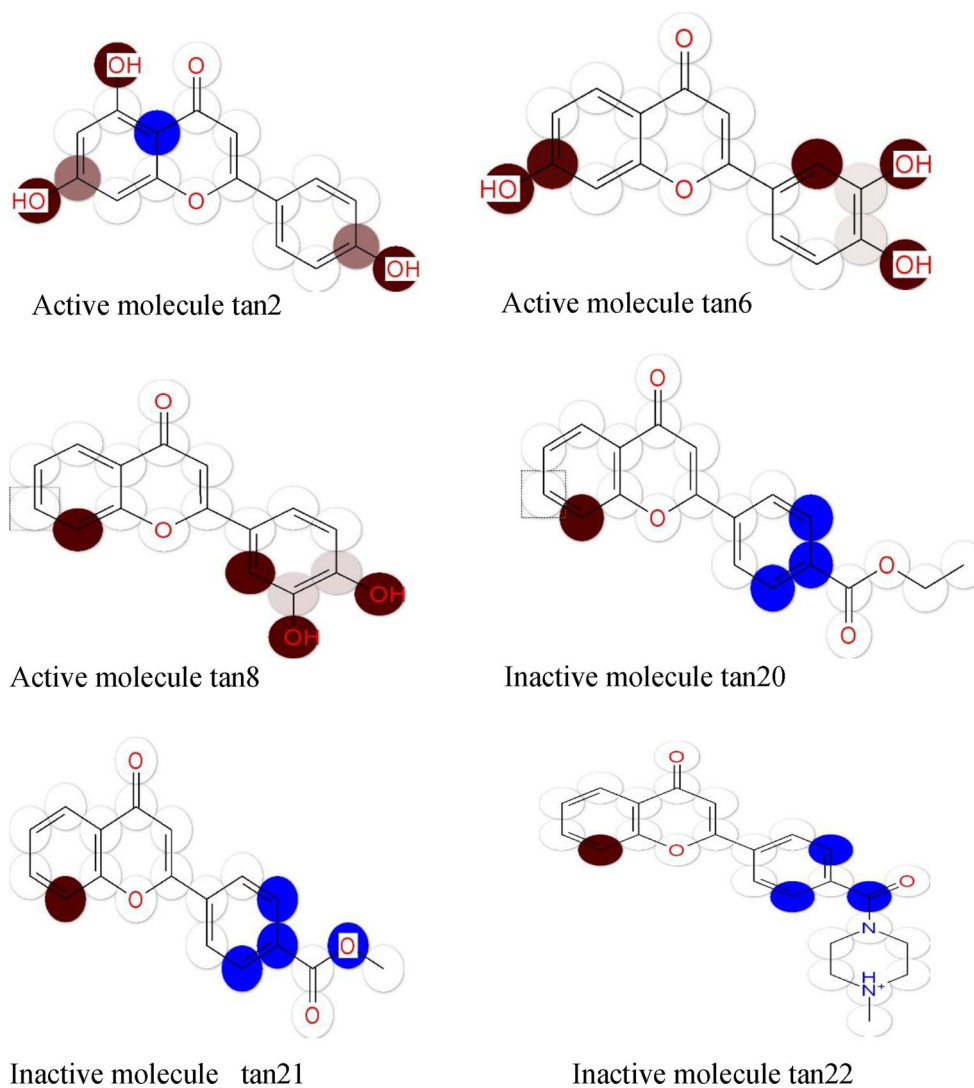
Active molecule tan6

Active molecule tan8

Inactive molecule tan20

Inactive molecule   tan21

Inactive molecule tan22

Fig. 16 Atomic contribution model for fingerprint—2D molprint

Table 13 KPLS results on training set

KPLS factors	SD	R^2
1	0.8353	0.6792
2	0.5433	0.8733
3	0.2874	0.9671

Table 14 KPLS results on test set

KPLS factors	RMS deviation	Q^2
1	0.7885	0.6307
2	0.9594	0.4533
3	0.9333	0.4790

Fig. 17 Actual versus predicted training set (fingerprint—fp dendritic)

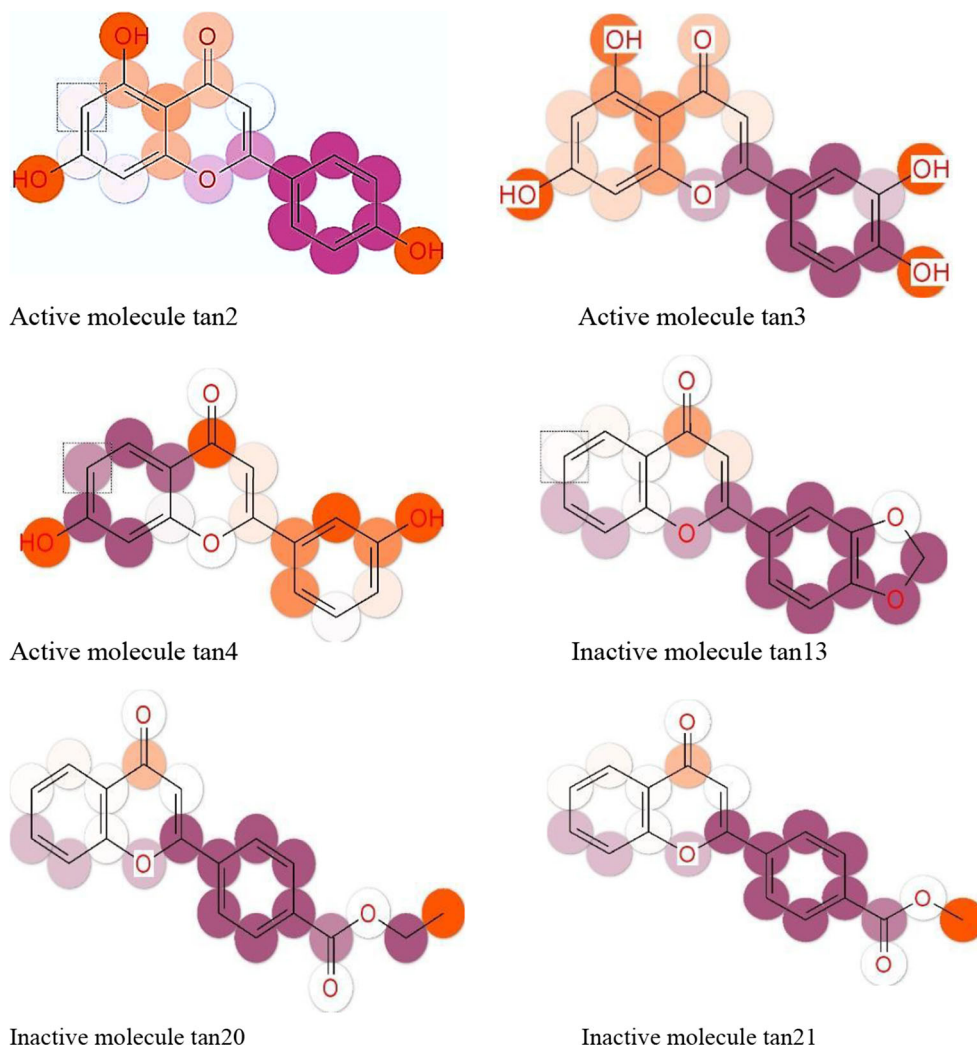**Fig. 18** Actual versus predicted test set (finger print—fp dendritic)

Active molecule tan2

Active molecule tan3

Active molecule tan4

Inactive molecule tan13

Inactive molecule tan20

Inactive molecule tan21
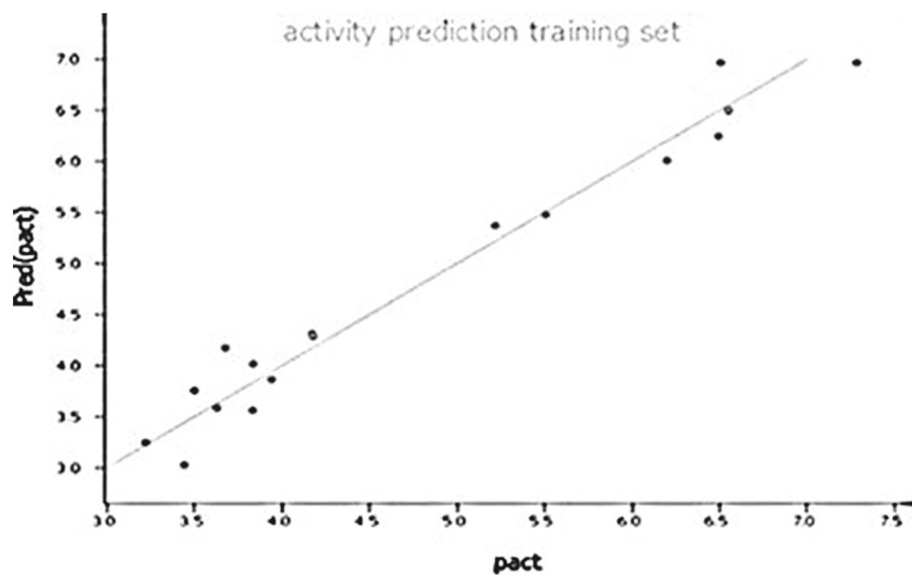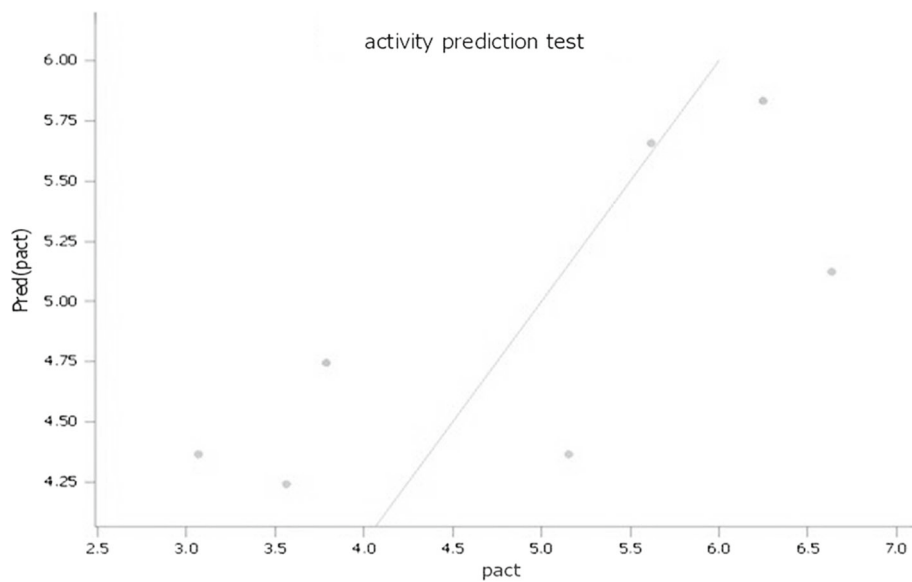
Fig. 19 Atomic contribution model for atom fingerprint—fp dendritic

Table 15 Activity prediction (actual vs. predicted)

Structure	Model set	Pact	Pred (pact)	Error
tan1	Training	6.495	6.242	− 0.253
tan2	Training	5.509	5.474	− 0.034
tan3	Test	5.62	5.655	0.036
tan4	Training	6.553	6.499	− 0.054
tan5	Test	6.252	5.832	− 0.42
tan6	Training	6.201	6.002	− 0.198
tan7	Training	6.509	6.976	0.468
tan8	Test	6.638	5.121	− 1.517
tan9	Training	6.553	6.484	− 0.069
tan10	Training	7.284	6.975	− 0.309

Table 15 continued

Structure	Model set	Pact	Pred (pact)	Error
tan11	Training	3.678	4.168	0.49
tan12	Training	3.225	3.24	0.015
tan13	Training	3.444	3.019	− 0.425
tan14	Training	3.633	3.58	− 0.052
tan15	Training	3.504	3.75	0.246
tan16	Test	3.071	4.364	1.293
tan17	Training	4.18	4.279	0.099
tan18	Training	5.222	5.363	0.141
tan19	Training	4.174	4.306	0.132
tan20	Test	3.565	4.239	0.673
tan21	Test	3.79	4.744	0.954
tan22	Training	3.836	3.554	− 0.281
tan23	Training	3.839	4.012	0.173
tan24	Test	5.155	4.364	− 0.791
tan25	Training	3.943	3.857	− 0.087

Fig. 20 Generation of Hash codes in dendritic fingerprint
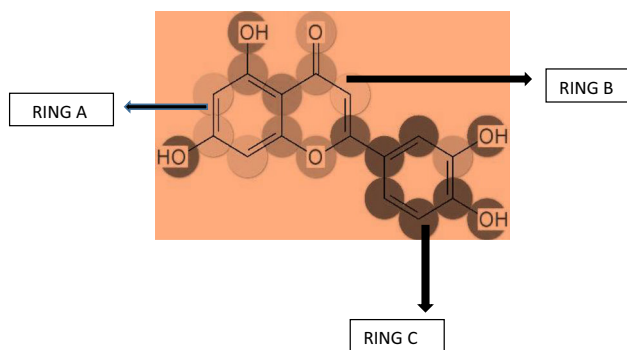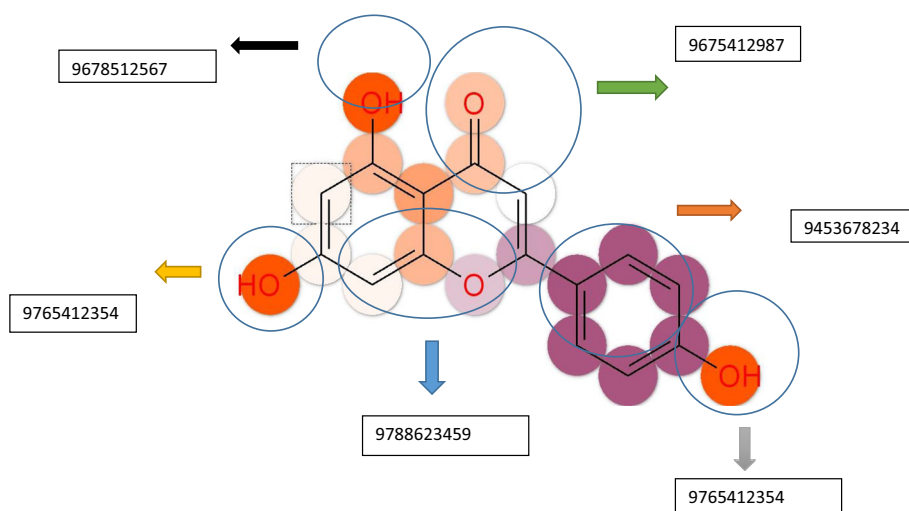

Fig. 21 Interpretation of QSAR model

Conclusions

Kernel-based partial least square regression was performed on a series of novel mono-substituted flavones using five binary fingerprinting methods. The contribution of each fingerprint model to the activity depends on several factors. In the present study, the fingerprint atom pairs gave a statistically significant 2D QSAR model with excellent regression values. The factors responsible for the success of pairwise fingerprint are molecular size and molecular weight. All the selected test compounds were bicyclic fused systems with mono-substitutions on rings A and C. The molecular weight of all the test compounds was in the range of 200–400 Dal-

tons. The fingerprint atom triplets also gave a statistically significant 2D QSAR equation due to the involvement of atom triplets which occur at particular site and at a particular distance. The three remaining fingerprints, namely linear, 2D Molprint and dendritic, could not reach acceptable regression values. This failure may be attributed to the atom typing scheme and structural variation. Suitable fingerprints should be selected based on structure, molecular size, types of ring systems and nature of extended branching.

References

1. Ferlay J, Shin HR, Bray F, Forman D, Mathers C, Parkin DM (2010) Estimates of worldwide burden of cancer in 2008: GLOBOCAN 2008. Int J Cancer 127:2893–2917. https://doi.org/10.1002/ijc.25516
2. Morin PJ, Sparks AB, Korinek V, Barker N, Clevers H, Vogelstein B, Kinzler KW (1997) Activation of β-catenin-TCF signaling in colon cancer by mutations in β-catenin or APC. Science 275:1787–1790. https://doi.org/10.1126/science.275.5307.1787
3. De Sousa EMF, Vermeulen L, Richel D, Medema JP (2011) Targeting WNT signaling in colon cancer stem cells. Clin Cancer Res 17:647–653. https://doi.org/10.1158/1078-0432.CCR-10-1204
4. Ikeda S, Kishida M, Matsuura Y, Usui H, Kikuchi A (2000) GSK-3 [beta]-dependent phosphorylation of adenomatous polyposis coli gene product can be modulated by [beta]-catenin and protein phosphatase 2A complexed with Axin. Oncogene 19:537. https://doi.org/10.1038/sj.onc.1203359
5. Huang SMA, Mishina YM, Liu S, Cheung A, Stegmeier F, Michaud GA, Hild M (2009) Tankyrase inhibition stabilizes axin and antagonizes Wnt signalling. Nature 461:614–620. https://doi.org/10.1038/nature08356
6. Donigian JR, de Lange T (2007) The role of the poly (ADP-ribose) polymerase tankyrase1 in telomere length control by the TRF1 component of the shelterin complex. J Biol Chem 282:22662–22667. https://doi.org/10.1074/jbc.M702620200
7. Narwal M, Koivunen J, Haikarainen T, Obaji E, Legala OE, Venkannagari H, Lehtiö L (2013) Discovery of tankyrase inhibiting flavones with increased potency and isoenzyme selectivity. J Med Chem 56:7880–7889. https://doi.org/10.1021/jm201510p
8. Schrödinger Release 2017-2: Maestro, Schrödinger, LLC, New York, NY (2017)
9. Okouchi S, Saegusa H (1989) Prediction of soil sorption coefficients of hydrophobic organic pollutants by adsorbability index. Bull Chem Soc Jpn 62:922–924. https://doi.org/10.1246/bcsj.62.922
10. Dancoff SM, Quastler H (1953) The information content and error rate of living things. In: Essays on the use of information theory in biology. University of Illinois Press, Urbana, p 263
11. Bertz SH (1981) The first general index of molecular complexity. J Am Chem Soc 103:3599–3601. https://doi.org/10.1021/ja00402a071
12. Ośmiałowski K, Halkiewicz J, Kaliszan R (1986) Quantum chemical parameters in correlation analysis of gas–liquid chromatographic retention indices of amines. J Chromatogr A 361:63–69
13. Karelson M, Lobanov VS, Katritzky AR (1996) Quantum-chemical descriptors in QSAR/QSPR studies. Chem Rev 496:1027–1040. https://doi.org/10.1021/cr950202r
14. Stanton DT, Jurs PC (1990) Development and use of charged partial surface area structural descriptors in computer-assisted quantitative structure–property relationship studies. Anal Chem 62:2323–2329. https://doi.org/10.1021/ac00220a013
15. Raychaudhury C, Ray SK, Ghosh JJ, Roy AB, Basak SC (1984) Discrimination of isomeric structures using information theoretic topological indices. J Comput Chem 5:581–588. https://doi.org/10.1002/jcc.540050612
16. Chenzhong C, Zhiliang L (1998) Molecular polarizability: a relationship to water solubility of alkanes and alcohols. J Chem Inf Comput Sci 38:1–7. https://doi.org/10.1021/ci9601729
17. Mulliken RS (1955) Electronic population analysis on LCAO–MO molecular wave functions. J Chem Phys 23:1833–1840. https://doi.org/10.1063/1.1740588
18. Knox JH, Kaliszan R (1985) Theory of solvent disturbance peaks and experimental determination of thermodynamic dead-volume in column liquid chromatography. J Chromatogr A 349:211–234
19. Luco JM, Yamin LJ, Ferretti HF (1995) Molecular topology and quantum chemical descriptors in the study of reversed-phase liquid chromatography:hydrogen-bonding behavior of chalcones and flavonones. J Pharm Sci 84:903–908. https://doi.org/10.1002/jps.2600840722
20. Hammette LP (1970) Physical organic chemistry: reaction rates, equilibria and mechanism. Mc Graw Hill, New York
21. Joshi RK, Meister T, Scapozza L, Ha TK (1994) A new quantum chemical approach in QSAR-analysis: parametrisation of conformational energies into molecular descriptors JMn (steric) and JSn (electronic). Arzneimittelforschung 44:779–790
22. Beckhaus HD (1978) $\mathcal{S}_f$ parameters: a measure of the front strain of alkyl groups. Angew Chem Int Ed Engl 17:593–594. https://doi.org/10.1002/cber.19781110107
23. Ivanciuc O, Balaban AT (1996) Design of topological indices: a new topological parameter for the steric effect of alkyl substituents. Croat Chem Acta 69:75–83. https://doi.org/10.1021/ci034266b
24. Dash SC, Behera GB (1980) A new steric parameter to explain ortho-substituent effect. Indian J Chem Sect A 19:541–543
25. Miyaki Y, Einaga Y, Fujita H (1978) Excluded-volume effects in dilute polymer solutions: very high molecular weight polystyrene in benzene and cyclohexane. Macromolecules 11:1180–1186. https://doi.org/10.1021/ma60066a022
26. Carbó R, Leyda L, Arnau M (1980) How similar is a molecule to another? An electron density measure of similarity between two molecular structures. Int J Quantum Chem 17:1185–1189
27. Parr RG, Pearson RG (1983) Absolute hardness: companion parameter to absolute electronegativity. J Am Chem Soc 105:7512–7516. https://doi.org/10.1021/ja00364a005
28. Kier LB (1989) An index of molecular flexibility from kappa shape attributes. Mol Inf 8:221–224. https://doi.org/10.1021/acs.jcim.6b00565
29. Sastry M, Lowrie JF, Dixon SL, Sherman W (2010) Large-scale systematic analysis of 2D fingerprint methods and parameters to improve virtual screening enrichments. J Chem Inf Model 50:771–784. https://doi.org/10.1021/ci100062n
30. Duan J, Dixon SL, Lowrie JF, Sherman W (2010) Analysis and comparison of 2D fingerprints: insights into database screening performance using eight fingerprint methods. J Mol Graph 29:157–170. https://doi.org/10.1016/j.jmgm.2010.05.008
31. Schrödinger Release 2016-3: Canvas, Schrödinger, LLC, New York, NY (2016)