



# Development and interpretation of a QSAR model for in vitro breast cancer (MCF-7) cytotoxicity of 2-phenylacrylonitriles

David T. Stanton<sup>1</sup> · Jennifer R. Baker<sup>2</sup> · Adam McCluskey<sup>2</sup> · Stefan Paula<sup>3</sup>

Received: 9 January 2021 / Accepted: 25 April 2021 / Published online: 4 May 2021  
© The Author(s), under exclusive licence to Springer Nature Switzerland AG 2021

## Abstract

The Arylhydrocarbon Receptor (AhR), a member of the Per-ARNT-SIM transcription factor family, has been as a potential new target to treat breast cancer sufferers. A series of 2-phenylacrylonitriles targeting AhR has been developed that have shown promising and selective activity against cancerous cell lines while sparing normal non-cancerous cells. A quantitative structure–activity relationship (QSAR) modeling approach was pursued in order to generate a predictive model for cytotoxicity to support ongoing synthetic activities and provide important structure-activity information for new structure design. Recent work conducted by us has identified a number of compounds that exhibited false positive cytotoxicity values in the standard MTT assay. This work describes a good quality model that not only predicts the activity of compounds in the MCF-7 breast cancer cell line, but was also able to identify structures that subsequently gave false positive values in the MTT assay by identifying compounds with aberrant biological behavior. This work not only allows the design of future breast cancer cytotoxic activity in vitro, but allows the avoidance of the synthesis of those compounds anticipated to result in anomalous cytotoxic behavior, greatly enhancing the design of such compounds.

**Keywords** QSAR · Model development · Model interpretation · Drug design · Breast cancer · MCF-7 · 2-phenylacrylonitriles · MTT assay

## Introduction

Breast cancer is the most common cancer in women, and the second-most lethal, worldwide. There is currently no cure for metastatic (also known as stage IV) breast cancer, and the 5-year survival rate for this invasive disease is approximately 25% [1]. Disease sufferers are categorized according to the hormone receptors expressed by their tumors. Breast cancer tumors are typically estrogen, progesterone or HER-2 receptor positive, and a number of targeted therapies are applicable to them [2]. Unfortunately, tumors expressing none of these receptors, known as labelled triple-negative tumors,

have no current targeted treatment [3]. With breast cancer incidence on the rise, and current treatments ineffective for advanced disease sufferers, new targeted treatments, with new biological targets, are urgently required [4, 5].

We and others have identified the Arylhydrocarbon Receptor (AhR) as a potential breast cancer drug target [6–8]. The AhR is a transcription factor member of the basic helix-loop-helix Per-ARNT-SIM family, and is traditionally associated with the metabolism of xenobiotic ligands [9, 10]. Many of the known AhR ligands, both exogenous and endogenous, and agonistic or antagonistic, are classed as either halogenated aromatic hydrocarbons (HAH) or poly-aromatic hydrocarbons (PAH) [11, 12]. Besides its role in mediating the toxicity of environmental contaminants, the AhR has also been associated with the possible treatment of diseases such as immune- and inflammatory-related conditions, inflammatory bowel disease, rheumatoid arthritis and multiple sclerosis (among others). It has recently been linked to some of the more insidious effects of the novel SARS-CoV-2 disease [13] as well as a possible pathway target for the treatment of Duchenne’s muscular dystrophy [14].

✉ David T. Stanton  
stantondt@gmail.com

<sup>1</sup> Belchertown, MA 01007, USA

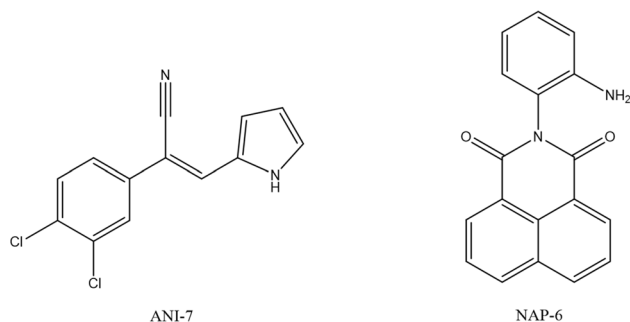
<sup>2</sup> Chemistry, School of Environmental & Life Sciences, The University of Newcastle, University Drive, Callaghan, NSW 2308, Australia

<sup>3</sup> Department of Chemistry, Tschannen Science Center, California State University at Sacramento, 6000 J Street, Sacramento, CA 95819, USA

Our interest commenced based on phenotypic observations that ANI-7 and NAP-6 (Fig. 1) showed high levels of breast cancer cell line specificity. Initial observations were made with the MCF-7 breast cancer cell line, with subsequent evaluations revealing high levels of activity against a wide array of drug resistant breast cancer cell lines. Biochemical investigations revealed the primary cell target as the AhR [15–17]. The development of new ligands targeting the AhR was subsequently conducted via *in silico* methods with a homology model of the ligand-binding domain [6, 18]. Efficacy and potency of the ligands was evaluated via a cytotoxic screen of 10 cancerous cell lines and one healthy breast cell line (MCF-10A) [19], utilizing an MTT [3-(4,5-dimethylthiazol-2-yl)-2,5-diphenyl tetrazolium bromide] assay [18]. Those ligands exhibiting potency and specificity towards the breast cancer cell lines were further investigated in a specific AhR reporter assay to verify their mechanism of action [15, 16].

Our recent efforts afforded a range of ligands that, at face value, exhibited exquisite selectivity to the breast cancer cell lines examined (up to 500-fold, over the healthy breast cell line) and excellent potency ( $GI_{50}$  as low as  $< 1$  nM). Subsequent morphological examination of the cells, however, demonstrated that the ligands were producing a metabolite that interfered with the tetrazolium moiety in MTT to form a formazan precipitate resulting in an aberrant MTT result. Rescreening of these compounds in an SRB (sulforhodamine B) assay [20], which does not rely on mitochondrial reductase enzymes, afforded a vastly different result [21].

Standard SAR evaluation of the MTT data and our homology model failed to provide an insight into the apparent increase in activity that we were noting, especially in the cases of the piperazine and piperidine analogues [21]. Docking of these analogues within the AhR binding pocket of the homology model provided no rational explanation for the observed MTT potency. Thus, we felt that this model was lacking in key descriptors of the ligand-AhR interaction. We decided to include a different modeling aspect in this work



**Fig. 1** Structures of ANI-7 and NAP-6, the initial ligands from our previous work that demonstrated breast cancer cytotoxicity via their action in the AhR pathway

in order to augment other structure design and data analysis methods used earlier. A quantitative structure–activity relationship, QSAR, approach was pursued because of the characteristics of the modeling method itself and the phenotypic nature of the biological activity. Unlike structure-based modeling methods which depend on identifying presumed target proteins and simulating ligand docking to the target, a QSAR approach makes no assumptions with regard to the biological target, and instead focuses on identifying differences in structural features of the ligands that correlate with differences in the observed biological activity. Since the biological activity is based on a whole-cell assay, some differences in activity may be related to properties other than just the binding and inhibition of the target protein such as cell wall permeation and other intermolecular interactions within the cytosol, and such aspects can be captured in a QSAR model. The physical interpretation of the QSAR model can provide a unique perspective on the SAR helping to identify structural features of compounds that convey activity that can be exploited for structure design. Such a model also provides the ability to rapidly predict the activity of a large number of candidate structures for synthesis planning, or virtual screening of large databases for candidate compounds as their demands on computational resources are modest and cost much less than physical screening. Lastly, the process of developing a QSAR model can be highly diagnostic of problems in the experimental data itself, and can identify compounds whose behavior differs from others with very similar structures.

## Experimental

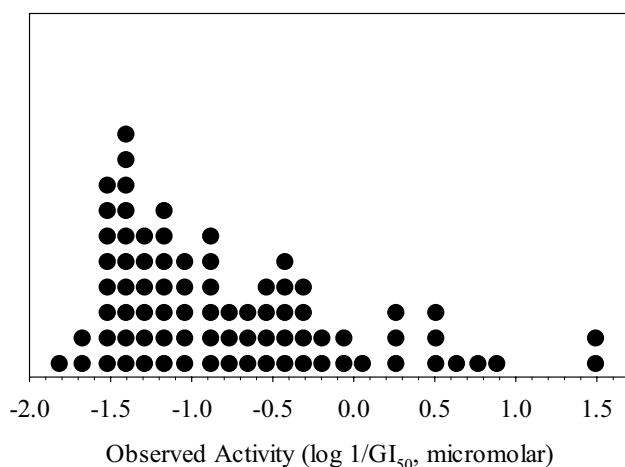
### Data

The chemical structures for 80 2-phenylacrylonitrile based AhR targeting compounds and their related biological data were collected from published reports [6, 19, 21–24]. The biological property of interest was growth inhibition,  $GI_{50}$ , defined as the concentration of the test compound yielding a 50% decrease in the growth of estrogen receptor positive (ER +ve) human breast cancer cells, MCF-7, relative to an untreated control. Structures for which no  $GI_{50}$  values were reported, or for which only single point assay values were reported (observed % inhibition at a single fixed concentration) were excluded from consideration. Since no difference in activity was detected for enantiomerically pure compounds [21], a single non-stereoisomer specific 2D structure was used to represent reported stereoisomers, and the structure was included only once in the analysis and the average activity of the isomers was assigned to that one structure. The original activity data were reported as micromolar ( $\mu$ M) concentrations and converted to  $\log(1/GI_{50})$  for modeling

purposes. The distribution of the observed activity data is shown in Fig. 2. The structures were drawn into the computer using MarvinSketch (ChemAxon, version 19.12.0) and stored in an Instant JChem database (ChemAxon, version 19.8.0). IUPAC names were obtained using MarvinSketch. The chemical names, structure labels, and observed biological data are provided in Table 1. The structures are provided in the form of Daylight SMILES strings and 2D structure diagrams in the Supplemental Information (Table S1 and S2, respectively).

## Model development

The structures were exported from Instant JChem as Daylight SMILES. Molecular structure descriptors were computed from SMILES using winMolconn (Hall Associates Consulting, version 1.0.2.1). These are topology-based (so-called “2D” descriptors) that do not require the generation of a three-dimensional conformation for a structure in order to compute the descriptor. They include molecular connectivity [25], electrotopological [26] and related descriptors. These topological descriptors were used because they help identify specific features of molecular structure and the structural environment surrounding them, and this information is required for extracting a detailed SAR while eliminating arbitrary decisions regarding the generation of relevant conformations [27]. A set of 1650 descriptors was computed for each of the 80 structures. These were filtered by removing any descriptor with greater than 90% zero values or greater than 90% identical non-zero values. A set of 296 descriptors remained after filtering for each of the structures. A list of the labels of the descriptors remaining after filtering is provided in the Supplemental Information (Table S3).



**Fig. 2** A dot plot illustrating the distribution of the observed activity data ( $\log 1/GI_{50}$ ) for the full dataset ( $N=80$ ). Each symbol represents a single observation (structure)

Both simulated annealing [28] and genetic algorithm [29] methods were used for variable selection to define subsets of the original 269 filtered molecular structure descriptors for multilinear regression evaluation. Potential models comprising 7–13 descriptors were evaluated based on internal validation statistical criteria; coefficient of multiple determination ( $R^2$ ) [30], root mean squared error(s) [30], partial-F tests [30], variance inflation factors [30], and leave-one-out cross-validation  $R^2$  ( $Q^2_{LOO}$ ) [31]. Partial least-squares (PLS) regression analysis (Minitab, Release-14) was used as a diagnostic for overfitting [32], and robust regression analysis [33] was used to detect and diagnose outlying observations. Y-variable (response) randomization [34] was used as a diagnostic for chance correlation. Following the selection of a final model, PLS was again used to extract the underlying structure-activity relationship encoded in the model [32, 35].

## Results and discussion

### QSAR model

Initial data analysis included all the relevant data from the sources described. However, it was soon discovered that several of the observations were exhibiting undue influence in the model equations obtained. When these outlying observations were excluded from the training set and their activity predicted using a model generated in their absence, the predicted activity values were much lower than the reported observed activities. At the same time, anomalies in the experimental results from the MTT assays were observed. It was discovered that many of these same compounds interfered with the MTT assay resulting in misleadingly high activity values [21]. Since the experimental and modeling analysis identified the same compounds, their structures were excluded from further modeling work (see Supplemental Information, Table S4). During subsequent model development, four additional compounds were identified that behaved as outliers in a fashion similar to the interfering compounds described above, supporting a decision to exclude them from the analysis. One additional structure was observed to be a major outlier for a different reason. Thus, a set of five observations were set aside from the original 80 structure data set. Their identity and an examination of the potential reasons for their behavior are described later.

A final model was obtained for the remaining 75 observations. This model included seven molecular descriptors, and yielded a good fit to the observed growth inhibition values ( $R^2=0.726$ ,  $Q^2_{LOO}=0.663$ ). The details of this model are provided in Table 2. A brief description of each of the seven descriptors is provided in Table 3. The correlation of the fitted and observed values for the training set is illustrated in the fit plot in Fig. 3. The internal validation statistics for the

**Table 1** Structure identification and observed cell growth inhibition values for the 80 structures used in this study

Structure label	Observed growth inhibition (GI <sub>50</sub> , μM)	Observed growth inhibition (log 1/GI <sub>50</sub> , μM)	IUPAC Name
A1	17	− 1.23E+00	(Z)-2-phenyl-3-(1H-pyrrol-2-yl)acrylonitrile
A2	15	− 1.18E+00	(Z)-2-(4-fluorophenyl)-3-(1H-pyrrol-2-yl)acrylonitrile
A3	4	− 6.02E−01	(Z)-2-(4-chlorophenyl)-3-(1H-pyrrol-2-yl)acrylonitrile
A5	0.56	2.52E−01	(Z)-2-(3,4-dichlorophenyl)-3-(1H-pyrrol-2-yl)acrylonitrile
A7	65	− 1.81E+00	2-(4-fluorophenyl)-3-(1H-pyrrol-2-yl)propanenitrile
A8	26	− 1.41E+00	2-(4-chlorophenyl)-3-(1H-pyrrol-2-yl)propanenitrile
A9	37	− 1.57E+00	2-(4-aminophenyl)-3-(1H-pyrrol-2-yl)propanenitrile
A11	49	− 1.69E+00	(Z)-2-(3,4-dichlorophenyl)hept-2-enenitrile
A22	2.2	− 3.42E−01	(Z)-2-(3,4-dichlorophenyl)-3-(p-tolyl)acrylonitrile
A23	1.5	− 1.76E−01	(Z)-2-(3,4-dichlorophenyl)-3-(naphthalen-2-yl)acrylonitrile
A25	6.5	− 8.13E−01	(Z)-2-(3,4-dichlorophenyl)-3-(4-fluorophenyl)acrylonitrile
A26	4.3	− 6.33E−01	(Z)-3-(4-chlorophenyl)-2-(3,4-dichlorophenyl)acrylonitrile
A27	16	− 1.20E+00	(Z)-3-(4-bromophenyl)-2-(3,4-dichlorophenyl)acrylonitrile
A28	0.13	8.86E−01	(Z)-2-(3,4-dichlorophenyl)-3-(4-nitrophenyl)acrylonitrile
A29	7.2	− 8.57E−01	(Z)-3-(3-chlorophenyl)-2-(3,4-dichlorophenyl)acrylonitrile
A30	23	− 1.36E+00	(Z)-2-(3,4-dichlorophenyl)-3-(4-hydroxyphenyl)acrylonitrile
A31	0.6	2.22E−01	(Z)-2-(3,4-dichlorophenyl)-3-(4-methoxyphenyl)acrylonitrile
A32	25	− 1.40E+00	(Z)-4-(2-cyano-2-(3,4-dichlorophenyl)vinyl)phenyl acetate
A33	15	− 1.18E+00	(Z)-2-(3,4-dichlorophenyl)-3-(pyridin-4-yl)acrylonitrile
A35	28	− 1.45E+00	(Z)-2-(3,4-dichlorophenyl)-3-(3,5-dihydroxyphenyl)acrylonitrile
B2	31	− 1.49E+00	(Z)-2-(3,4-dichlorophenyl)-3-(4-(trifluoromethyl)phenyl)acrylonitrile
B3	1.3	− 1.14E−01	(Z)-2-(3,4-dichlorophenyl)-3-(1H-indol-3-yl)acrylonitrile
B4	7	− 8.45E−01	(Z)-2-(3,4-dichlorophenyl)-3-(furan-2-yl)acrylonitrile
B14	2.8	− 4.47E−01	(Z)-2-(3,4-dichlorophenyl)-3-(2-methyl-1H-indol-3-yl)acrylonitrile
B15	3.7	− 5.68E−01	(Z)-2-(3,4-dichlorophenyl)-3-(5-methyl-1H-indol-3-yl)acrylonitrile
B16	2.3	− 3.62E−01	(Z)-3-(5-chloro-1H-indol-3-yl)-2-(3,4-dichlorophenyl)acrylonitrile
B17	6.9	− 8.39E−01	(Z)-3-(5-bromo-1H-indol-3-yl)-2-(3,4-dichlorophenyl)acrylonitrile
B19	4	− 6.02E−01	(Z)-3-(1H-benzo[g]indol-3-yl)-2-(3,4-dichlorophenyl)acrylonitrile
B20	0.23	6.38E−01	(Z)-2-(3,4-dichlorophenyl)-3-(1H-indol-5-yl)acrylonitrile
C1	27	− 1.43E+00	(E)-3-(4-chlorophenyl)-2-(1H-pyrrole-2-carbonyl)acrylonitrile
C11	23	− 1.36E+00	(E)-3-(perfluorophenyl)-2-(1H-pyrrole-2-carbonyl)acrylonitrile
C18	35	− 1.54E+00	(E)-2-(2H-isoindeole-1-carbonyl)-3-(4-nitrophenyl)acrylonitrile
C21	13	− 1.11E+00	(2E)-2-[(E)-1H-indole-3-carbonyl]-3-(2,3,4,5,6-pentafluorophenyl)prop-2-enenitrile
C23	11	− 1.04E+00	3-(1H-indol-3-yl)-2-[(1H-indol-3-yl)methyl]-3-oxopropanenitrile
C27	34	− 1.53E+00	(E)-2-cyano-N-(4-methoxybenzyl)-3-(1H-pyrrol-2-yl)acrylamide
C28	6	− 7.78E−01	(E)-2-cyano-N-(3,4-dichlorobenzyl)-3-(1H-pyrrol-2-yl)acrylamide
C29	27	− 1.43E+00	(E)-2-cyano-3-(furan-2-yl)-N-(4-methoxybenzyl)acrylamide
C30	20	− 1.30E+00	(E)-2-cyano-N-(3,4-dichlorobenzyl)-3-(furan-2-yl)acrylamide
C31	21	− 1.32E+00	(E)-2-cyano-N-(4-methoxybenzyl)-3-(5-methylfuran-2-yl)acrylamide
C33	9	− 9.54E−01	(E)-3-(5-chlorofuran-2-yl)-2-cyano-N-(4-methoxybenzyl)acrylamide
C34	11	− 1.04E+00	(E)-3-(5-bromofuran-2-yl)-2-cyano-N-(4-methoxybenzyl)acrylamide
C35	3	− 4.77E−01	(E)-2-cyano-N-(4-methoxybenzyl)-3-(5-phenylfuran-2-yl)acrylamide
C36	20	− 1.30E+00	(E)-3-(4-bromofuran-2-yl)-2-cyano-N-(4-methoxybenzyl)acrylamide
C37	7	− 8.45E−01	(E)-2-cyano-3-(furan-3-yl)-N-(4-methoxybenzyl)acrylamide
C39	18	− 1.26E+00	(E)-2-cyano-N-(4-methoxybenzyl)-3-phenylacrylamide
C40	36	− 1.56E+00	(E)-2-cyano-N-(4-methoxybenzyl)-3-(p-tolyl)acrylamide
C41	11	− 1.04E+00	(E)-3-(4-chlorophenyl)-2-cyano-N-(4-methoxybenzyl)acrylamide

**Table 1** (continued)

Structure label	Observed growth inhibition (GI <sub>50</sub> , μM)	Observed growth inhibition (log 1/GI <sub>50</sub> , μM)	IUPAC Name
C44	33	− 1.52E+00	( <i>E</i> )-2-cyano-N-(4-methoxybenzyl)-3-(naphthalen-2-yl)acrylamide
C45	16	− 1.20E+00	( <i>E</i> )-2-cyano-N-(4-methoxybenzyl)-3-(naphthalen-1-yl)acrylamide
C46	29	− 1.46E+00	( <i>E</i> )-2-cyano-N-(3,4-dichlorobenzyl)-3-phenylacrylamide
C48	21	− 1.32E+00	( <i>E</i> )-3-(4-chlorophenyl)-2-cyano-N-(3,4-dichlorobenzyl)acrylamide
C50	45	− 1.65E+00	( <i>E</i> )-2-cyano-N-(3,4-dichlorobenzyl)-3-(4-methoxyphenyl)acrylamide
C51	29	− 1.46E+00	( <i>E</i> )-2-cyano-N-(3,4-dichlorobenzyl)-3-(naphthalen-2-yl)acrylamide
C52	8	− 9.03E−01	( <i>E</i> )-2-cyano-N-(3,4-dichlorobenzyl)-3-(naphthalen-1-yl)acrylamide
D6	2.5	− 3.98E−01	( <i>Z</i> )-2-(4-bromophenyl)-3-(1 <i>H</i> -pyrrol-2-yl)acrylonitrile
D7	15	− 1.18E+00	( <i>Z</i> )-2-(2-fluorophenyl)-3-(1 <i>H</i> -pyrrol-2-yl)prop-2-enenitrile
D8	9.8	− 9.91E−01	( <i>Z</i> )-2-(3-fluorophenyl)-3-(1 <i>H</i> -pyrrol-2-yl)acrylonitrile
D10	1.7	− 2.30E−01	( <i>Z</i> )-3-(1 <i>H</i> -pyrrol-2-yl)-2-(4-(trifluoromethyl)phenyl)acrylonitrile
D12	1.9	− 2.79E−01	( <i>Z</i> )-2-(3-chlorophenyl)-3-(1 <i>H</i> -pyrrol-2-yl)acrylonitrile
D14	23	− 1.36E+00	( <i>Z</i> )-2-(2,4-dichlorophenyl)-3-(1 <i>H</i> -pyrrol-2-yl)acrylonitrile
E13	26	− 1.41E+00	( <i>Z</i> )-2-(2,6-dichlorophenyl)-3-(2-nitrophenyl)acrylonitrile
E14	25	− 1.40E+00	( <i>Z</i> )-2-(2,6-dichlorophenyl)-3-(3-nitrophenyl)acrylonitrile
E15	2.9	− 4.62E−01	( <i>Z</i> )-2-(3,4-dichlorophenyl)-3-(2-nitrophenyl)acrylonitrile
E16	5.3	− 7.24E−01	( <i>Z</i> )-2-(3,4-dichlorophenyl)-3-(3-nitrophenyl)acrylonitrile
E18	0.89	5.06E−02	( <i>Z</i> )-3-(5-bromo-1 <i>H</i> -pyrrol-2-yl)-2-(3,4-dichlorophenyl)acrylonitrile
E19	1	0.00E+00	( <i>Z</i> )-3-(4-bromo-1 <i>H</i> -pyrrol-2-yl)-2-(3,4-dichlorophenyl)acrylonitrile
E20	0.33	4.81E−01	( <i>Z</i> )-3-(4,5-dibromo-1 <i>H</i> -pyrrol-2-yl)-2-(3,4-dichlorophenyl)acrylonitrile
E21	0.48	3.19E−01	( <i>Z</i> )-2-(3,4-dichlorophenyl)-3-(3,4,5-tribromo-1 <i>H</i> -pyrrol-2-yl)acrylonitrile
E26	3.5	− 5.44E−01	( <i>Z</i> )-2-(2,6-dichloro-3-nitrophenyl)-3-(2-nitrophenyl)acrylonitrile
E27	12	− 1.08E+00	( <i>Z</i> )-2-(2,6-dichloro-3-nitrophenyl)-3-(3-nitrophenyl)acrylonitrile
E28	7.4	− 8.69E−01	( <i>Z</i> )-2-(2,6-dichloro-3-nitrophenyl)-3-(4-nitrophenyl)acrylonitrile
E29	2.8	− 4.47E−01	( <i>Z</i> )-2-(3-amino-2,6-dichlorophenyl)-3-(2-aminophenyl)acrylonitrile
E30	4.5	− 6.53E−01	( <i>Z</i> )-2-(3-amino-2,6-dichlorophenyl)-3-(3-aminophenyl)acrylonitrile
E32	0.32	4.95E−01	( <i>Z</i> )-3-(4-bromo-3-nitrophenyl)-2-(3,4-dichlorophenyl)acrylonitrile
E35	0.03	1.52E+00	( <i>Z</i> )-3-(4-aminophenyl)-2-(3,4-dichlorophenyl)acrylonitrile
E36	0.17	7.70E−01	( <i>Z</i> )-2-(3,4-dichlorophenyl)-3-(4-(methylamino)phenyl)acrylonitrile
E37	0.28	5.53E−01	( <i>Z</i> )-2-(3,4-dichlorophenyl)-3-(4-(dimethylamino)phenyl)acrylonitrile
E38	0.034	1.47E+00	( <i>Z</i> )-N-(4-(2-cyano-2-(3,4-dichlorophenyl)vinyl)phenyl)acetamide
E39	2.1	− 3.22E−01	( <i>Z</i> )-3-(1 <i>H</i> -benzo[d]imidazol-6-yl)-2-(3,4-dichlorophenyl)acrylonitrile
F13c	13	− 1.11E+00	( <i>Z</i> )-3-(4-{3-[4-(4-chlorophenyl)amino]-2-hydroxypropoxy}phenyl)-2-(3,4-dichlorophenyl)prop-2-enenitrile

The structure label indicates the source of the data; the letter identifies the published article and the number is the number assigned to the structure in that article. The letter assignments are A—Tarleton, et al. [19], B—Tarelton, et al. [22], C—Tarleton, et al. [23], D—Al Otaibi, et al. [24], E—Baker, et al. [6], F—Baker, et al. [21]

model are good. The partial-F values are all greater than the critical value of 2.15 (F-distribution with 7 and 67 degrees of freedom,  $\alpha = 0.05$ ), indicating that each descriptor is significant in the model given the presence of the other six. The variance inflation factors are generally small (below 10.0), and average 4.46 indicating that there is minimal collinearity between the descriptors. PLS analysis validates seven components, indicating a lack of over fitting. Valid components are those for which the *Wold criterion* holds. This states that inclusion of components into the model terminates when the

ratio of the PRedicted Error Sum of Squares (PRESS) values for the *i*th and the *i*th − 1 component exceeds 1.0 [36]. This corresponds to the point where  $Q^2_{LOO}$  for the *i*th component decreases from the  $Q^2_{LOO}$  of the *i*th − 1 component.

Since a large number of molecular descriptors (X-variables) were analyzed in variable selection step ( $N_X = 296$ ) relative to the number of observations ( $N_{obs} = 75$ ), it is possible that the model is a result of a random correlation [37]. While this may be true for a variable space of these dimensions composed of totally random and uncorrelated

**Table 2** Details of the QSAR model derived for the set of 75 observations from the 2-phenylacrylonitrile data set

Descriptor label	Coefficient	Partial-F	Variance inflation factor
<i>n2pag13</i>	0.3754	81.24	5.42
<i>SdssC</i>	1.606	93.97	4.52
<i>SaaCH</i>	0.2088	37.20	9.15
<i>xch5</i>	6.628	66.90	1.83
<i>netype22</i>	- 0.2500	25.86	3.72
<i>SssNH</i>	0.2865	31.09	2.55
<i>dxp9</i>	- 1.341	13.57	4.08
<i>y-intercept</i>	- 4.182		

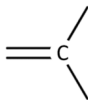
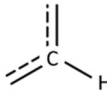
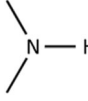

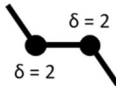
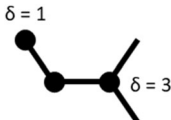
$R^2 = 0.726$ ,  $s = 0.344$ ,  $Q^2_{\text{LOO}} = 0.663$ , Overall F-value = 25.39

Equation form: Obs.  $\log(1/GI_{50}) = 1.606 \times SdssC + 0.2088 \times SaaCH + 0.2865 \times SssNH + 6.628 \times xch5 - 1.341 \times dxp9 - 0.2500 \times netype22 + 0.3754 \times n2pag13 - 4.182$

variables, the actual dimensionality of the descriptor spaces used in QSAR methods as practiced herein is actually much smaller because of a high degree of correlation between the molecular descriptors themselves [38]. For example,

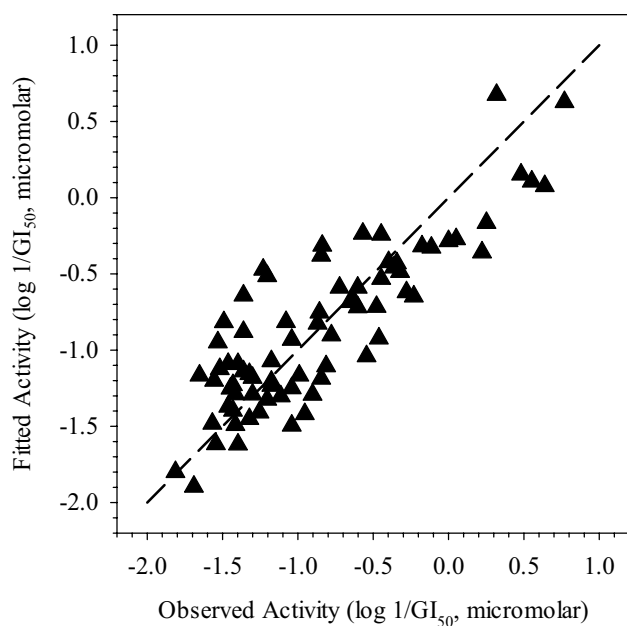
principal components analysis [39] of the 296 descriptors for the 75 observations analyzed in this instance showed that 27 principal components explain 99.0% of the variance in the descriptor space. So, the underlying dimensionality of the system is much smaller than it at first appears making it much less likely that the model is the result of a chance correlation. Nonetheless, the method of Y-randomization was used to evaluate the possibility of chance correlation. This was done by randomly scrambling the order of the Y variable (observed  $GI_{50}$ ) values without changing the ordering of the X variables (molecular descriptors). A set of ten random Y variables were generated from the original Y variable using Minitab-14. The maximum absolute value of the Pearson correlation coefficient [40] for the original Y variable and the randomized Y variables was 0.196 (minimum = 0.00437, average = 0.0738). Each of the 10 randomized response variables was used to generate new 7-variable regression equations using both the genetic algorithm and simulated annealing methods yielding 20 final models. The average  $R^2$  for these 20 models was 0.387 (minimum = 0.305, maximum = 0.483), and the average  $Q^2_{\text{LOO}}$  value was 0.242 (minimum = 0.121, maximum = 0.376). One should note that hundreds of models are evaluated for each

**Table 3** Labels, a brief description and a diagram of the key molecular structure feature for each of the seven molecular descriptors in the final model

Descriptor label	Description	Key structural feature
SdssC	Sum of the atom level E-State of all carbon atoms in the molecule of type =C < [26]	
SaaCH	Sum of the atom level E-State of all unsubstituted aromatic carbon atoms in the molecule [26]	
SssNH	Sum of the atom level E-State of all nitrogen atoms in the molecule of type -NH- [26]	
xch5	Simple 5th-order chain (five ring bonds) molecular connectivity index (Only the ring bonds are considered in this version, no extra-ring bonds are included in the subgraph) [25]	
dxp9	Simple 9th-order path difference molecular connectivity index. Computed by taking the difference between xp9 for the structure in question and the same descriptor for the hypothetical unbranched version of the structure with the same atom count and atom types [25]	
netype22	Count of single edges between two delta-2 vertices	
n2pag13	Count of 2nd-order path subgraphs (two consecutive bonds) between a delta-1 vertex and a delta-3 vertex	

The descriptor labels within the winMolconn output are case sensitive





**Fig. 3** Fit plot for the QSAR model showing the correlation of the fitted and observed activity values for the 75 observations in the model training set

single run of model generation when using these variable selection methods, and that far more than 20 total equations were considered in generating these results. Since the Y-randomized model results are substantially poorer than those obtained for the original Y variable model ( $R^2=0.726$ ,  $Q^2_{LOO}=0.663$ ), we conclude that the model is not a result of chance correlation.

### Structure activity analysis of the model

In addition to providing a means to predict the activity of new structures, the model can be analyzed to extract

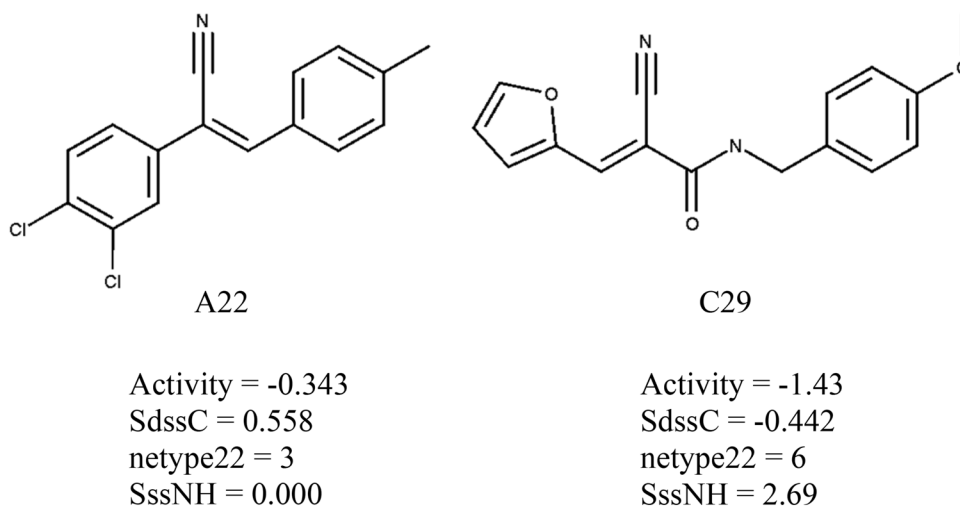
information that explains how specific changes in structure affect the observed activity of the compound. Such information provides a greater understanding of the underlying SAR, and can provide detailed information for designing new synthetic candidates. A method for conducting such an analysis using PLS analysis has been described previously [32, 35]. This method was used in this work. The details of conducting the analysis are not provided here, but a summary of the findings for each PLS component is provided.

As noted above, PLS analysis validates all seven components of the model, and each component explains a part of the underlying SAR in the model. The model explains 72.6% of the variance in the observed data. A summary of the findings from each component is provided below. In this discussion, the signed squared PLS weights for highly weighted descriptors from each component are given in parentheses following the descriptor label.

#### Component-1

The first component provides 36.3% of the information in the model. Three of the descriptors are highly weighted in this component; SdssC (+ 52.0%), netype22 (− 29.6%), and SssNH (− 15.1%). Due to the non-uniform distribution of the observed activity data, the model uses this component to help explain the bulk of the data with observed activity in the range of − 1.5 and − 0.3 (log 1/GI<sub>50</sub> values). This component is focused on the linking group containing the nitrile and forms the connection between the terminal ring systems. Examples of a more active structure (A22) and a less active structure (C29) are shown in Fig. 4. More active structures have only the favorable ethylene linker, where less active structures include an amide group and an additional methylene carbon atom in the linker. The SdssC descriptor has positive values for the more active structures. The

**Fig. 4** Structures and molecular descriptor values for examples illustrating the key structure features identified in Component-1 of the model. Activity value is log 1/GI<sub>50</sub>, micromolar

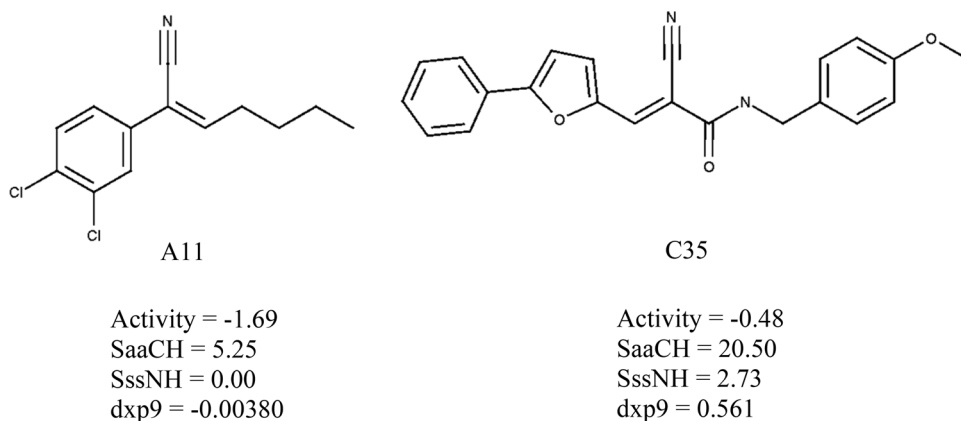


less active structures take a negative value for SdssC due to the presence of the carbonyl group. Additionally, the less active structures have positive non-zero values for the SssNH descriptor which captures information regarding the amide nitrogen atom. Combined with a negative weight the SssNH descriptor contributes to fit lower observed activity values. The amide nitrogen atom is missing from the more favorable ethylene linker, which is reflected in a value of zero for the SssNH descriptor for the more active structures. The presence of the combination of the methylene carbon atom and the amide nitrogen atom in the linker is captured by a larger value for the ntype22 descriptor, which is combined with a negative weight and further contributes to fitting the lower activity observations.

### Component-2

Component-2 provides an additional 16.3% of the information in the model (52.6% cumulative) and shows three highly weighted descriptors; SaaCH (+28.4%), SssNH (+25.7%), and dxp9 (+24.1%). This component focuses on the influence of the terminal portions of the structure for a particular subset of molecules. Figure 5 provides examples of structures which highlight the features that are the focus of which Component-2. SAR corrections are made to structures containing the disfavored linker identified in Component-1 which included the amide group and methylene carbon (see C35, Fig. 5). Those structures are identified using the SssNH descriptor which takes a positive weight in Component-2. The descriptors SaaCH and dxp9 descriptors, both taking positive weights, work together to highlight additional and larger aromatic ring systems present in several molecules with that linker which do not exhibit activity as low as suggested based on Component-1. This suggests that the presence of the unfavorable linker can be mitigated to a certain extent by the greater size of the terminal groups.

**Fig. 5** Structures and molecular descriptor values for examples illustrating the key structure features identified in Component-2 of the model. Activity value is  $\log 1/GI_{50}$ , micromolar



### Component-3

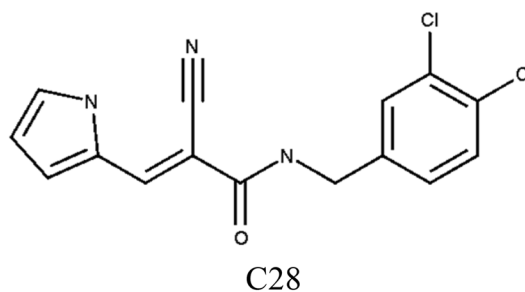
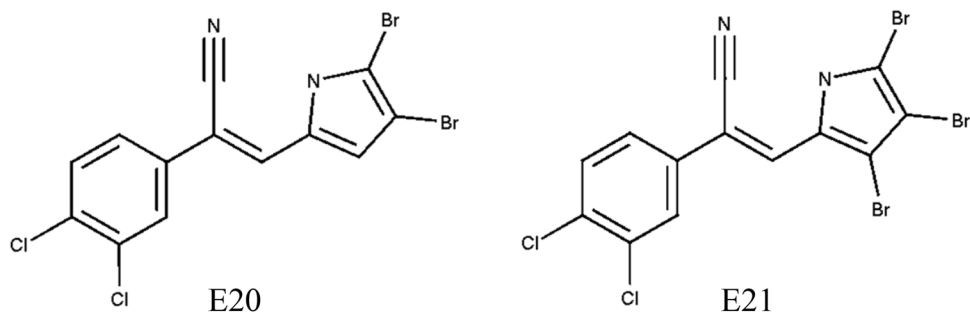
Component-3 makes further corrections to the SAR described in Component-1. This component expresses an additional 21.1% of the information in the model (63.7% cumulative) based on four highly weighted descriptors; n2pag13 (+28.3%), dxp9 (-23.4%), SssNH (+15.2%), and xch5 (+11.5%). Component-3 primarily acts to identify structure features which appear to increase the activity over that expressed in Component-1. Structures illustrating these features are shown in Fig. 6. The n2pag13 descriptor captures the presence of the substituents on the rings of the terminal ends of the structures, particularly the pyrrole ring which is captured by the xch5 descriptor (see E20 and E21, Fig. 6). A small up-correction is also indicated for the inclusion of a pyrrole ring for C28 which contains the unfavorable linker group identified in Component-1 and is captured by the positively weighted SssNH descriptor in Component-3. Combined, these features suggest that smaller rings, and in particular those with large halogen substituents, can be used to replace larger ring systems to achieve increased activity, even given the presence of other unfavorable features.

### Component-4

Component-4 accounts for an additional 9.07% of the information in the model (82.8% cumulative), and involves three highly weighted descriptors; dxp9 (-44.1%), SdssC (+25.2%), and SaaCH (+12.5%). This component primarily makes corrections downward in activity from the SAR expressed in the previous components. Example structures illustrating the key features are shown in Fig. 7. Component-4 focuses primarily on the number and nature of the substituents on the rings at the terminal ends of the molecule. The dxp9 and SaaCH descriptors combine to identify structures with the shorter more favorable linker, but which have either too many substituents, or less favorable substituents. The nitro groups (see E27 and E28) in particular appear to be disfavored as they reduce the value



**Fig. 6** Structures and molecular descriptor values for examples illustrating the key structure features identified in Component-3 of the model. Activity value is  $\log 1/GI_{50}$ , micromolar



Activity = -0.778  
n2pag13 = 4  
d xp9 = 0.134  
SssNH = 2.66  
xch5 = 0.144

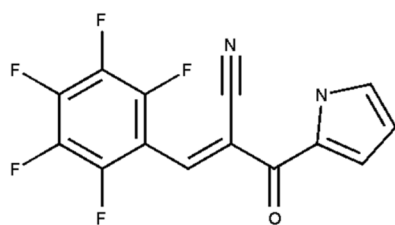
of the SaaCH descriptor much more than a simple halogen substituent does at the same position. The compounds C11 and C21 contain the pentafluoro benzene ring, and are also differentiated from other short linker molecules by the presence of the carbonyl group in the linker which is captured by the SdssC descriptor. Unlike the methylamide substructure of the longer linker identified in Component-1, these structures have only the additional carbonyl group which is why they were not fit properly in Component-1. The larger indole ring in C21 mitigates the impact of the multiple aromatic substituents and the carbonyl in the linker, but only to a relatively small extent. The much larger 1*H*-benzo[*g*]indole ring of B19 increases the value SaaCH, but significantly increases the value of d xp9 which is a more highly weighted change and suggests

the presence of the greater bulk at that end of the molecule is disfavored.

### Component-5

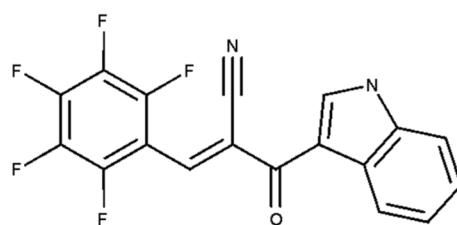
Component-5 explains an additional 10.0% of the SAR expressed in the model (92.8% cumulative), and it does this using three highly weighted descriptors: n2pag13 (+50.2%), xch5 (−36.5%), and SaaCH (+7.40%). Example structures related to Component-5 are shown in Fig. 8. The xch5 descriptor divides the data set into two clusters. One cluster generally takes a value of zero for xch5 indicating the absence of a five-membered ring whereas the other cluster takes a non-zero value for the xch5 descriptor indicating the presence of a five-membered ring, typically in the form of a pyrrole or indole. One structural feature in particular is the missing double bond in the linking group

**Fig. 7** Structures and molecular descriptor values for examples illustrating the key structure features identified in Component-4 of the model. Activity value is  $\log 1/GI_{50}$ , micromolar



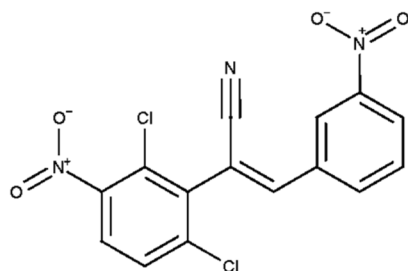
C11

Activity = -1.362  
 dxp9 = 0.0812  
 SdssC = -1.734  
 SaaCH = 4.08



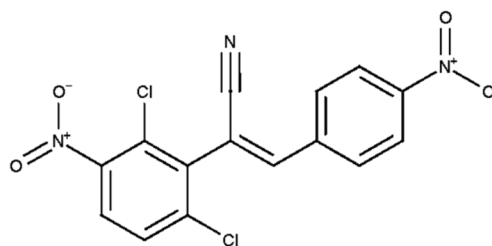
C21

Activity = -1.11  
 dxp9 = 0.406  
 SdssC = -1.70  
 SaaCH = 7.85



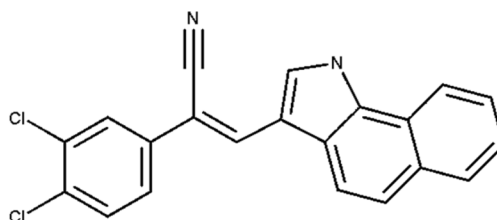
E27

Activity = -1.08  
 dxp9 = 0.353  
 SdssC = -0.0470  
 SaaCH = 7.96



E28

Activity = -0.869  
 dxp9 = 0.403  
 SdssC = 0.00240  
 SaaCH = 7.86



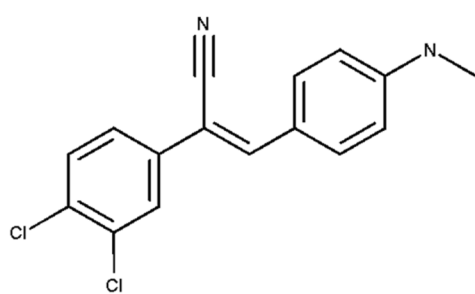
B19

Activity = -0.602  
 dxp9 = 1.29  
 SdssC = 0.535  
 SaaCH = 19.5

of several structures (A7, A8, and A9). These structures were overestimated in Component-1. The missing double bond affects the value of the xch5 and SaaCH which helps to identify the difference in the linking feature. Taken together, this component adjusts the activity of these structures down from what was captured in earlier

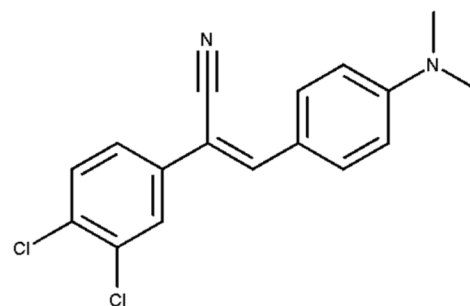
components. The n2pag13 descriptor makes its greatest contribution to the model in Component-5. Within the clusters, the n2page13 descriptor helps to capture structure features that impart greater activity than expressed in the previous components. This component puts special emphasis on branched substituents such as methylamine or dimethylamine (E36 and E37). The activity of these

**Fig. 8** Structures and molecular descriptor values for examples illustrating the key structure features identified in Component-5 of the model. Activity value is  $\log 1/GI_{50}$ , micromolar



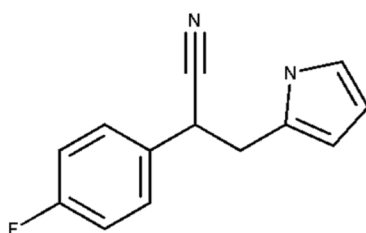
E36

Activity = 0.770  
n2pag13 = 4  
xch5 = 0.00



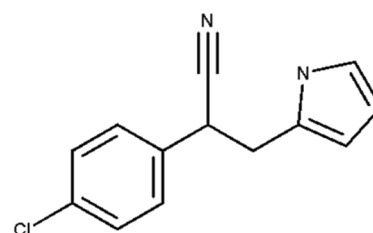
E37

Activity = 0.553  
n2pag13 = 5  
xch5 = 0.00



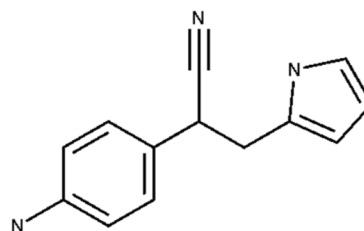
A7

Activity = -1.81  
n2pag13 = 1  
xch5 = 0.144



A8

Activity = -1.42  
n2pag13 = 1  
xch5 = 0.144



A9

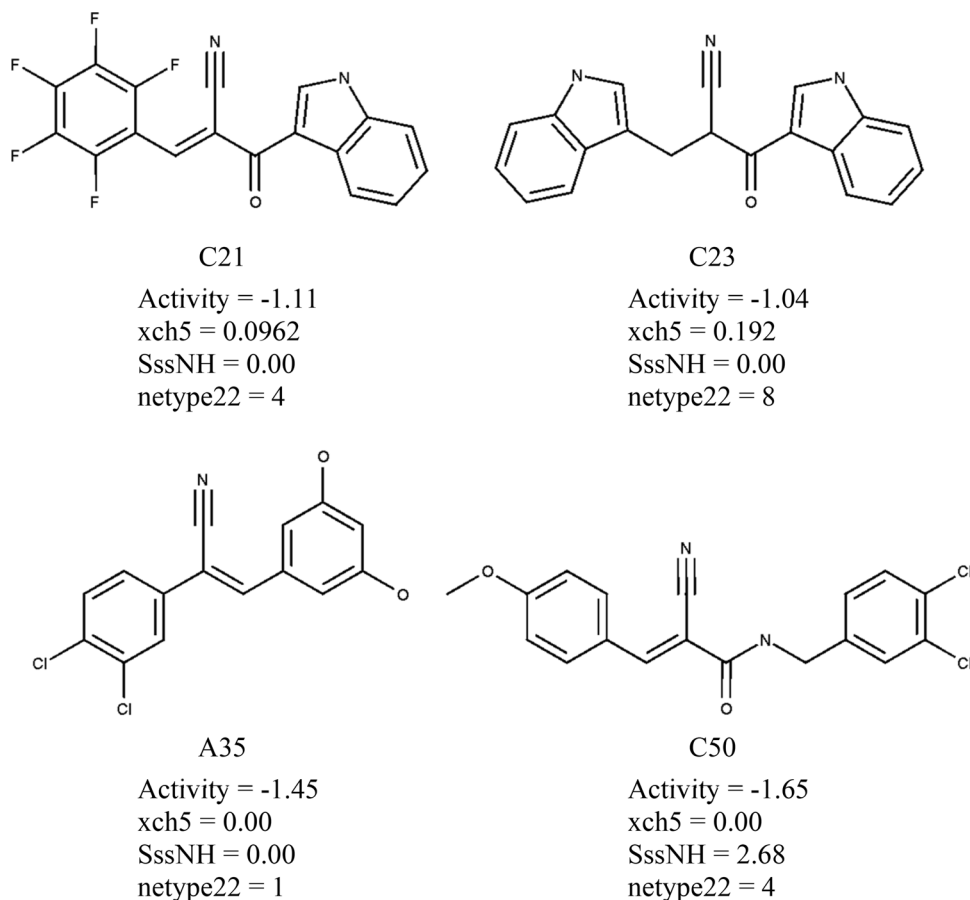
Activity = -1.57  
n2pag13 = 1  
xch5 = 0.144

structures is underestimated in the earlier components. The slightly more complex substituents affect the activity differently than adding simple single atom substituents such as halogens.

### Component-6

This component accounts for an additional 5.39% of the work done by the model (98.2% cumulative), and involves three highly weighted descriptors: xch5 (+40.5%), SssNH (-39.3%) and netype22 (+10.2%). Example structures that illustrate the important features for the component are shown in Fig. 9. The small contribution of the component

**Fig. 9** Structures and molecular descriptor values for examples illustrating the key structure features identified in Component-6 of the model. Activity value is  $\log 1/GI_{50}$ , micromolar



to the overall model indicates that it is accounting for only small corrections in a few specific structures. For example, small positive adjustments are made to fit the activity of C21 and C23 are captured by the combination of information from the xch5 descriptor (encodes information for the 5-membered ring), and netype22 which captures the additional unsubstituted aromatic carbon atoms in the indole ring systems. These structures were not as well fitted by the combination of earlier components. Similarly, a small correction is made in the fitting of the activity of A35 using netype22. There are additional substituents on this molecule that are more polar and hydrophilic than other related structures which are apparently less favored. The importance of the SssNH descriptor is in identifying a difference in C50 compared to other similar structures based on its inclusion of the slightly more hydrophilic methoxy substituent. The overall effect of this component is to provide small adjustments to the fitted activity of structures with unique rings and ring substituents.

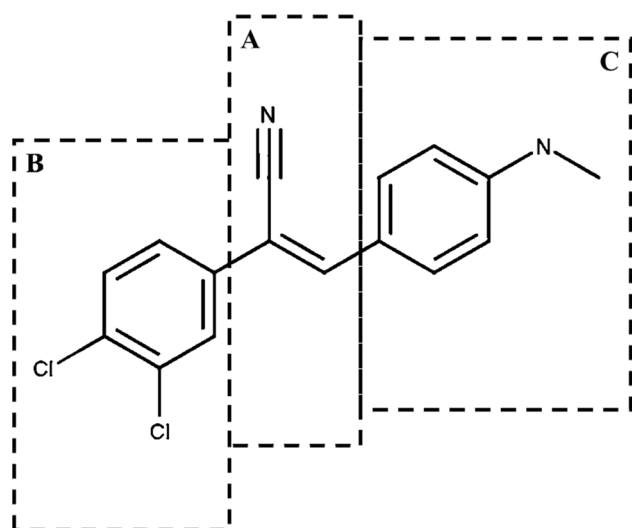
#### Component-7

This last component accounts for only an additional 1.8% of the model (100% cumulative) and uses three highly weighted

descriptors: netype22 (− 39.5%), SaaCH (+ 35.9%), and SdssC (− 16.1%). The purpose of this component is to make a correction to one unusual structure, A11 (see Fig. 5). The value of SaaCH is 5.25, which is nearly the lowest value of any structure in the data set. The value of SdssC is 0.659, which is one of the largest values in the data set. Combined with the netype22 descriptor (value = 4) and the signs of the respective PLS weights, these descriptors capture the presence of the aliphatic chain and the absence of the second ring system present in the rest of the data set, which contributes to making this the analog with the lowest activity in the data set.

#### SAR analysis summary

The PLS analysis provides very detailed information regarding the underlying SAR represented in the model. Figure 10 summarizes the extracted SAR in graphic form. The model addresses structure variations by dividing the structure space into three regions. There are two main themes in the SAR; the nature of the central linking group, and the nature of the terminal groups. Overall, the nature of the linking group makes the largest SAR contribution (see Component-1), and there are generally two classes of linker (Fig. 10, region A).



**Fig. 10** A graphic illustration of the important structure regions identified in the structure activity relationship extracted using PLS analysis. Two main features were identified; The linking group, (a), connecting the two terminal groups, (b) and (c). The structure for E36 is used to illustrate features favored in each region

A short linker group is preferred, with the ethylene linkage being most favored. Longer linking groups including either a carbonyl or a combination of amide and a methylene carbon are disfavored. The key feature is not just length, but the presence of the double bond seems to be an important required feature. Structures differing only by the absence of the double bond can exhibit a decrease in activity from single digit and low double digit micromolar activity to mid-double digit micromolar activity (see Component-5). The positive contribution of the unsaturation in the linker may be related to the electrostatics of the connection between the terminal moieties, or may provide a desirable conformational constraint.

Within two linker subsets, the nature of the terminal groups is quite diverse and that diversity plays the role of enhancing or diminishing the activity within a linker class. The model size (number of descriptors) is a direct result

of the diversity of the terminal groups. Less diversity was experimentally explored in region B, but some specific trends have been observed. It is clear from the PLS analysis that the 3,4-dichlorophenyl group is most favored fragment present in region B (see Fig. 10). A number of other substituents and substituent patterns were evaluated, but none were as active overall. For example, 2,4-dichloro and 2,5-dichloro analogs are not as active as the 3,4-dichloro substitution pattern. Other substitution patterns that approach similar size and orientation are 4-bromo and 4-trifluoromethyl and exhibit similar activity. This suggests that ring systems that put greater bulk in that region would be favored, and favoring hydrophobic bulk in particular. While adding bulk in that region, a polar nitro substituent is disfavored (see Component-4), probably because while not being hydrophilic, it is not found to be as hydrophobic as other substituent types [41].

Greater diversity has been experimentally explored in region C. A preference for moderate-sized groups in this region (see Fig. 10) is also clear from the PLS analysis. Phenyl and pyrrole rings with hydrophobic substituents and unsubstituted indole rings are favored in region C, as is illustrated in PLS components 2, 3, 5 and 6. Component-6 indicates hydrophilic substituents are disfavored. PLS component-7 shows that a lack of sufficient bulk in region C is also disfavored.

Due to the nature of PLS, a significant advantage of this type of SAR analysis is that the SAR trends described are orthogonal. The results suggest that each region can be investigated and optimized more or less separately. This can simplify planning of new synthetic targets to generate more detail regarding the scope of what is possible in each region as well as increasing what can be achieved with a given reaction route.

## Review of outliers

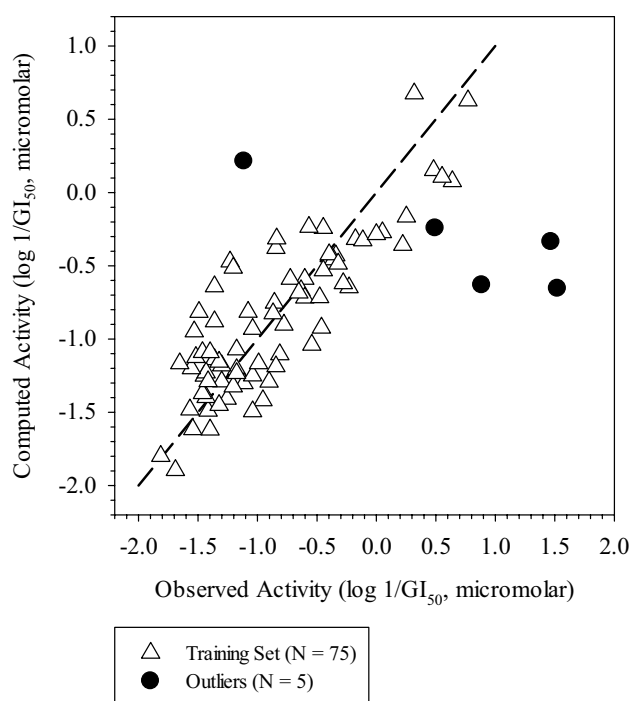
An examination of outliers can be instructive because they can help to detect either experimental issues (assay anomalies, sample identity or purity issues, etc.) or identify

**Table 4** Identity, observed and predicted activity (MTT, log 1/GI<sub>50</sub>, micromolar) for the five outliers identified during model development

Structure label	Observed activity (MTT, log 1/GI <sub>50</sub> , μM)	Predicted activity (MTT, log 1/GI <sub>50</sub> , μM)	Prediction error	Observed activity (SRB, log 1/GI <sub>50</sub> , μM)
A28	0.886	- 0.631	1.52	- 1.26
E32	0.495	- 0.242	0.737	- 0.398
E35	1.52	- 0.655	2.18	- 0.380
E38	1.47	- 0.335	1.80	ND
F13c	- 1.11	0.214	- 1.33	ND

The available results from the SRB assay are also shown

ND not determined



**Fig. 11** Predicted activity results for the five modeling outliers computed using the model. The prediction results are shown in the context of the training set fitted results

modeling or data analysis issues (data processing problems, descriptor selection limitations, structure diversity issues, etc.). During the development of the model, a set of five statistical outliers were detected and set aside. The outliers are identified in Table 4. The model was used to predict the activity values for the outliers. A comparison of the predicted and observed activity values for the outliers are shown graphically in Fig. 11.

Examination of the predicted activity values for the outliers based on the model shows that the activity of four of the five structures is underpredicted; the observed activity is much greater than the predicted activity. In general, one would expect a random error to be more evenly distributed, so this pattern suggests a systematic problem and raises possibility that these large positive errors are suspicious. This is the same behavior that was detected in preliminary modeling work, and that was found to be related to interference caused by some compounds in the MTT assay. In work done after model development was already under way, experimental evaluation of activity of compounds A28, E32 and E35 using the SRB assay verified that they were, in fact, false positives in the MTT assay due to the compounds interfering with the method. The overall SAR for the model suggests hydrophobic substituents are favored in region C of the structure (Fig. 10). The observed activity of E38 is greater than expected given the more hydrophilic substituent in that region of

the structure. For example, A30 has a hydroxyl substituent in the para-position and has an observed activity value of  $-1.36$ . A26 has a chlorine in the same position and has an activity value of  $-0.633$ . While it is possible that E38 represents a real departure from the overall SAR captured by the model, based on past observations of interfering compounds in the MTT assay and on the comparison to relatively similar compounds, its exclusion from the training set seems reasonable pending the acquisition of additional experimental data.

The remaining outlier, F13c, is larger than any of the structures in the training set. The most unique part of the structure resides in region C (Fig. 10) of the structure where there are no structures with similarly large features. This structure is quite dissimilar from the structures in training set in this respect which takes it out of the knowledge domain of the model and makes it a structural outlier. It is reasonable to set it aside until other structures which similarly expand into region C are experimentally evaluated. This will be explored in future work.

## Conclusions

The AhR, a transcription factor, has been identified as a potential good druggable target for the future treatment of breast cancer and that 2-phenylacrylonitriles likely act against this receptor. In this work, a good quality model was obtained for the cytotoxicity of a series of 2-phenylacrylonitriles and related compounds against MCF-7 human breast cancer cells. The model provides the means to predict activity of compounds not yet synthesized and also provides a detailed description of the underlying SAR which can be used to guide subsequent rounds of structure design. Additionally, recent experimental efforts have identified a number of compounds that gave inaccurate results in the standard MTT assay, with analogues expressing potency manyfold more active than they were later determined to be. The model development process showed it was possible to identify structures that departed significantly from the SAR expressed by the majority of the available data, which was agreement with the newly obtained experimental results. The model has identified several other observations likely to be interfering in the assay in the same way, which will need to be verified experimentally. This work enables design of future analogues targeting the AhR pathway which will be reported in due course.

**Supplementary Information** The online version contains supplementary material available at <https://doi.org/10.1007/s10822-021-00387-5>.



**Acknowledgements** The authors wish to express their appreciation to the reviewer who suggested addressing the possibility of chance correlation in the development of the model.

**Funding** Jennifer R. Baker University of Newcastle Postgraduate Research Scholarship, Adam McCluskey University of Newcastle Priority Research Centre for Drug Development.

**Data availability** All data used in this work is included in either the tables of the main manuscript, or in the associated Supplementary information.

**Code availability** Not applicable.

## Declarations

**Conflict of interest** None.

## References

- Krohe M, Hao Y, Lamourex RE, Galipeau N, Globe D, Foley C, Mazar I, Solomon J, Shields AL (2016) Patient-reported outcomes in metastatic breast cancer: a review of industry—sponsored clinical trials. *Breast Cancer Basic Clin Res* 10:93–102
- Greenwalt I, Zaza N, Das S, Li BD (2019) Precision medicine and targeted therapies in breast cancer. *Surg Oncol Clin N Am* 29:51–62
- Tang Y, Wang Y, Kiani MF, Wang B (2016) Classification, treatment strategy, and associated drug resistance in breast cancer. *Clin Breast Cancer* 16:335–343
- Fares J, Kanojia D, Cordero A, Rashidi A, Miska J, Schwartz CW, Savchuk S, Ahmed AU, Balyasnikova IV, Cristofanilli M, Gradishar WJ, Lesniak MS (2019) Current state of clinical trials in breast cancer brain metastases. *Neuro Oncol Pract* 6:392–401
- Nedeljković M, Damjanović A (2019) Mechanisms of chemotherapy resistance in triple-negative breast cancer: how we can rise to the challenge. *Cells* 8:957
- Baker JR, Gilbert J, Paula S, Zhu X, Sakoff JA, McCluskey A (2018) Dichlorophenylacrylonitriles as AhR ligands that display selective breast cancer cytotoxicity in vitro. *ChemMedChem* 14:1447–1458
- Powell JB, Goode GD, Eltom SE (2014) The aryl hydrocarbon receptor: a target for breast cancer therapy. *J Cancer Ther* 4:1177–1186
- Baker JR, Sakoff JA, McCluskey A (2020) The arylhydrocarbon receptor (AhR) as a breast cancer drug target. *Med Res Rev* 40:972–1001
- Okey AB, Bondy GP, Mason ME, Nebert DW, Forster-Gibson CJ, Muncan J, Dufresne MJ (1980) Temperature-dependent cytosol-nucleus translocation of the Ah receptor for 2,3,7,8-tetrachlorodibenzo-p-dioxin in continuous cell culture lines. *J Biol Chem* 255:11415–11422
- Stockinger B, Di Meglio P, Gialitakis M, Duarte JH (2014) The aryl hydrocarbon receptor: multitasking in the immune system. *Annu Rev Immunol* 32:403–432
- Denison MS, Nagy SR (2003) Activation of the aryl hydrocarbon receptor by structurally diverse exogenous and endogenous chemicals. *Annu Rev Pharmacol Toxicol* 43:309–334
- Tarnow P, Tralau T, Luch A (2019) Chemical activation of estrogen and arylhydrocarbon receptor signalling pathways and their interaction in toxicology and metabolism. *Expert Opin Drug Metab Toxicol* 15:219–229
- Turski WA, Wnorowski A, Turski AN, Turski CA, Turski L (2020) AhR and IDO1 in pathogenesis of Covid-19 and the “Systemic AhR Activation Syndrome:” a translational review and therapeutic perspectives. *Restor Neurol Neurosci* 38:343–354
- Wilkinson IVL, Perkins KJ, Dugdale H, Moir L, Vuorinen A, Chatzopoulou M, Squire SE, Monecke S, Lomov A, Geese M, Charles PD, Burch P, Tinsley JM, Wynne GM, Davies SG, Wilson FX, Rastinejad F, Mohammed S, Davies KE, Russell AJ (2020) Chemical proteomics and phenotypic profiling identifies the aryl hydrocarbon receptor as a molecular target of utrophin modulator ezutromid. *Angew Chem Int Ed* 59:2420–2428
- Gilbert J, De Iulius GN, Tarleton M, McCluskey A, Sakoff JA (2018) (Z)-2-(3,4-Dichlorophenyl)-3-(1H-Pyrrol-2-yl)acrylonitrile exhibits selective antitumour activity in breast cancer cell lines via the Aryl hydrocarbon Receptor pathway. *Mol Pharmacol* 93:168–177
- Gilbert J, De Iulius GN, McCluskey A, Sakoff JA (2020) A novel naphthalimide that selectively targets breast cancer via the arylhydrocarbon receptor pathway. *Sci Rep* 10:13978
- Baker JR, Pollard BL, Lin AJS, Gilbert S, Paula S, Zhu X, Sakoff JA, McCluskey A (2021) Modelling and phenotypic screening of NAP-6 and 10-Cl-BBQ AhR ligands displaying selective breast cancer cytotoxicity in vitro. *ChemMedChem* (In press). <https://doi.org/10.1002/cmdc.202000721>
- Paula S, Baker JR, Zhu X, McCluskey A (2019) Binding of chlorinated phenylacrylonitriles to the aryl hydrocarbon receptor: computational docking and molecular dynamics simulations. In: Stefaniu A (ed) *Molecular docking and molecular dynamics*. Rijeka, Intech
- Tarleton M, Gilbert J, Robertson MJ, McCluskey A, Sakoff JA (2011) Library synthesis and cytotoxicity of a family of 2-phenylacrylonitriles and discovery of an estrogen dependent cancer lead compound. *Med Chem Commun* 2:31–37
- Skehan P, Storeng R, Scudiero D, Monks A, McMahon J, Vistica D, Warren JT, Bokesch H, Kenney S, Boyd MR (1990) New colorimetric cytotoxicity assay for anticancer-drug screening. *J Natl Cancer Inst* 82:1107–1112
- Baker JR, Russell CC, Gilbert J, Sakoff JA, McCluskey A (2020) Amino alcohol acrylonitriles as activators of the aryl hydrocarbon receptor pathway, an unexpected MTT phenotypic screening outcome. *ChemMedChem* 15:490–505
- Tarleton M, Gilbert J, Sakoff JA, McCluskey A (2012) Cytotoxic 2-phenylacrylonitriles, the importance of the cyanide moiety and discovery of potent broad spectrum cytotoxic agents. *Eur J Med Chem* 57:65–73
- Tarleton M, Dyson L, Gilbert J, Sakoff JA, McCluskey A (2013) Focused library development of 2-phenylacrylamides as broad spectrum cytotoxic agents. *Bioorgan Med Chem* 21:333–347
- Al Otaibi A, Gordon CP, Gilbert J, Sakoff JA, McCluskey A (2014) The influence of ionic liquids on the Knoevenagel condensation of 1H-pyrrole-2-carbaldehyde with phenyl acetonitriles-cytotoxic 3-substituted-(1H-pyrrol-2-yl)acrylonitriles. *RSC Adv*. <https://doi.org/10.1039/c3ra47418f>
- Kier LB, Hall LH (1986) *Molecular connectivity in structure-activity analysis*. Research Studies Press, Letchworth
- Kier LB, Hall LH (1999) *Molecular structure description. The electrotopological state*. Academic Press, San Diego
- Stanton DT (2008) On the importance of topological descriptors in understanding structure-property relationships. *J Comput Aided Mol Des* 22:441–460
- Sutter JM, Jurs PC (1995) Selection of molecular descriptors for quantitative structure-activity relationships. *Data Handl Sci Tech* 15:111–132

29. Luke BT (1996). In: Devillers J (ed) Genetic algorithms in molecular modeling. Academic Press, New York, pp 35–66
30. Kutner MH, Nachtsheim CJ, Neter J, Li W (2005) Applied linear statistical models, 5th edn. McGraw-Hill Irwin, New York
31. Suhuurmann G, Ebert R, Chen J, Wang B, Kuhne R (2008) External validation and prediction employing the predictive squared correlation coefficient—Test set activity mean vs training set activity mean. *J Chem Inf Model* 48:2140–2145
32. Stanton DT (2012) QSAR and QSPR model interpretation using partial least squares (PLS) analysis. *Curr Comput Aided Drug Des* 8:107–127
33. Rousseeuw PJ, Leroy AM (1987) Robust regression & outlier detection. Wiley, New York
34. Wold S, Eriksson L (1995). In: van de Waterbeemd H (ed) Statistical validation of QSAR results. VCH, New York, pp 309–318
35. Stanton DT (2003) On the physical interpretation of QSAR models. *J Chem Inf Comput Sci* 43:1423–1433
36. Wold S (1978) Cross-validatory estimation of the number of components in factor and principal components models. *Technometrics* 20:397–405
37. Topliss JG, Edwards RP (1979) Chance factors in studies of quantitative structure-activity relationships. *J Med Chem* 22:1238–1244
38. Katritzky AR, Dobchev DA, Slavov S, Karelson M (2008) Legitimate utilization of large descriptor pools for QSAR/QSPR models. *J Chem Inf Model* 48:2207–2213
39. Johnson RA, Wichern DW (1988) Applied multivariate statistical analysis, 2nd edn. Prentice-Hall, Englewood Cliffs, New Jersey, p 341
40. Ott L (1988) An introduction to statistical methods and data analysis, 3rd edn. PWS-Kent, Boston, pp 319–323
41. Sagawa N, Shikata T (2014) Are all polar molecules hydrophilic? Hydration numbers of nitro compounds and nitriles in aqueous solution. *Phys Chem Chem Phys* 16:13262–13270

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.