ERpred: A web server for the prediction of subtype-specific estrogen receptor antagonists

Nalini Schaduangrat, Aijaz Ahmad Malik, Chanin Nantasenamat*

Center of Data Mining and Biomedical Informatics, Faculty of Medical Technology, Mahidol University, Bangkok 10700, Thailand

*Corresponding Author:
Chanin Nantasenamat
Center of Data Mining and Biomedical Informatics, Faculty of Medical Technology, Mahidol University, Bangkok 10700, Thailand
Email address: chanin.nan@mahidol.edu

Abstract

Estrogen receptor alpha and beta (ERα and ERβ) are responsible for breast cancer metastasis through their involvement of clinical outcomes. Estradiol and hormone replacement therapy targets both ERs, but this often leads to an increased risk of breast and endometrial cancers as well as thromboembolism. A major challenge is posed for the development of compounds possessing ER subtype specificity. Herein, we present a large-scale classification structure-activity relationship (CSAR) study of inhibitors from the ChEMBL database which consisted of an initial set of 11,618 compounds for ERα and 7,810 compounds for ERβ. The IC_{50} was selected as the bioactivity unit for further investigation and after the data curation process, this led to a final data set of 1,593 and 1,287 compounds for ERα and ERβ, respectively. We employed the random forest (RF) algorithm for model building and of the 12 fingerprint types, models built using the PubChem fingerprint was the most robust (Ac of 94.65% and 92.25% and Matthews correlation coefficient (MCC) of 89% and 76% for ERα and ERβ, respectively) and therefore selected for feature interpretation. Results indicated the importance of features pertaining to aromatic rings, nitrogen-containing functional groups and aliphatic hydrocarbons. Finally, the model was deployed as the publicly available web server called ERpred at http://codes.bio/erpred where users can submit SMILES notation as the input query for prediction of the bioactivity against ERα and ERβ.

36   **Keywords: breast cancer, estrogen, estrogen receptor, data science, quantitative**
37   **structure-activity relationship**

Introduction

41   Breast cancer is the most frequently detected cancer amongst women with over 2 million new

42   cases and an estimated 627,000 deaths (15% of all cancer deaths in women) in 2018, according to

43   the WHO (World Health Organization, 2018). Furthermore, it is a well-known fact that levels of

44   Estrogen Receptor (ER) impacting breast cancer metastasis are the fundamental and critical

45   determinants of clinical outcomes (Kammerer et al., 2013; Gamucci et al., 2013). In addition, ER

46   positive breast cancer types exhibit favorable responses to hormone therapy (Althuis et al., 2004;

47   Foulkes, Smith & Reis-Filho, 2010; Thrane et al., 2013), for example tamoxifen (Ramirez-Ardila

48   et al., 2013), or aromatase inhibitors (Hiscox, Davies & Barrett-Lee, 2009), designed to block

49   aberrant signaling within oncogenic pathways. However, a major obstacle in the case of

50   chemotherapy in ER-positive breast cancers is chemoresistance (Kim et al., 2010; Ji et al., 2019;

51   Han et al., 2019). Therefore, new systemic therapies are urgently needed.

52

53   Estrogen receptor is a member of the nuclear receptor family, mainly found in the nucleus but can

54   also be seen in the cytoplasm and mitochondria. ER consists of two main subtypes (i.e., ERα and

55   ERβ) which bind to hormones and trigger the activation or repression of genes (Brzozowski et al.,

56   1997). Estrogen signaling is selectively stimulated or inhibited depending on the balance between

57   ERα and ERβ activities in target organs. Both receptor subtypes are expressed in various cells and

58   tissues (i.e., breast, prostate and ovary) as they control various physiological functions of the

59   human body (i.e., reproductive, skeletal, cardiovascular and central nervous systems). The

60   mammary gland, uterus, ovary (thecal cells), bone, male reproductive organs, prostate, liver and

61   adipose tissues are mainly composed of ERα (Welboren et al., 2009). In contrast, ERβ is found

62   mainly in the prostate, bladder, ovary (granulosa cells), colon, adipose tissue, and the immune

63   system (Weiser, Foradori & Handa, 2008). Common physiological roles for ERα and ERβ includes

64   the development and function of ovaries and the protection of the cardiovascular system (Paterni

65   et. al., 2014). A more prominent role is exerted on the mammary gland and uterus as well as on

66   the homeostasis of the skeletal system and the regulation of metabolism by ERα. ERβ on the other

67   hand, exerts a more powerful effect on the central nervous and immune systems (Paterni et. al.,
68   2014). Furthermore, the β subtype is shown to generally counteract the hyperproliferation of ERα-
69   promoted cells in tissues such as the breast and uterus (Heldring et al., 2007).
70

71   Estradiol and hormone replacement therapy targets both ERs, but this often leads to an increased
72   risk of breast and endometrial cancers as well as thromboembolism. Selective estrogen receptor
73   modulators (SERMs) are the most common drug group used in ER-positive breast cancer treatment
74   with tamoxifen as the first line agent used to block mitogenic effects of estrogen at all stages of
75   breast cancer, particularly in pre- and post-menopausal patients (Abdulkareem & Zurmi, 2012). In
76   addition, Fulvestrant, the main drug in the group of Selective estrogen receptor downregulators
77   (SERDs) is used as an alternative in tamoxifen resistant breast cancers and acts by disrupting the
78   ER receptor and blocking ER dimerization which in turn inhibits estrogen signaling via ER down-
79   regulation (Osborne, Wakeling & Nicholson, 2004). However, the effectiveness of fulvestrant is
80   decreased by acquired resistance whereby a response to therapy is not seen in most ER-positive
81   breast cancer patients (Cook, Shajahan & Clarke, 2011). Furthermore, through mechanisms
82   distinct from ER subtype selective binding, SERMs target both receptor subtypes even though they
83   display tissue-selective agonist/antagonist activities (Paterni et. al., 2014). An ideal SERM would
84   thus possess antagonist activity in the mammary gland and uterus and antagonist activity in other
85   tissues such as those pertaining to the skeletal, cardiovascular or the central nervous systems
86   (Jordan, 2001). Alternatively, based on the distribution and levels of the two ER subtypes in the
87   various tissues mentioned above, subtype-selective ligands could be used to elicit beneficial
88   estrogen-like activities and reduce side effects. In this regard, there appears to be particular promise
89   for the use of subtype-selective agonists/antagonists (Kumar et al., 2011). However, the challenge
90   for developing ER subtype specific compounds remains elusive.
91

92   The ER structure contains a globular ligand binding domain (LBD) harboring a hormone-binding
93   site, a homo- or heterodimerization interface, and coregulator (activator and repressor) interaction
94   sites (Kumar et al., 2011). The amino acid sequence of ERα and ERβ displays a 55% sequence
95   identity in their respective LBDs, which represents a significant difference (Kerdivel, Habauzit &
96   Pakdel, 2013). The LBD of both ER subtypes are comprised of 12 α-helices (H1-H12) arranged in

128 and chemical properties. Such models are also useful for rationalizing the importance and

129 contributions of molecular features on investigated activities/properties. Subsequently, crucial

130 information pertaining to ER binding affinity coupled with structure-binding and structure-activity

131 relationship data, have led to the formulation of reliable ERα models (Anstead, Carlson &

132 Katzenellenbogen, 1997; Serafimova et al., 2007; Xiang et al., 2009; Toropov et al., 2012; Chang

133 et al., 2013; Ribay et al., 2016; Suvannang et al., 2017; Lee & Barron, 2017; Pavlin et al., 2018;

134 Pang et al., 2018; Balabin & Judson, 2018; Cotterill et al., 2019; Bafna et al., 2020). However, not

135 much research has been conducted regarding the binding specificity towards ERβ (Manas et al.,

136 2004; Coriano et al., 2018). Thus, the aim of this study is to build classification models able to (i)

137 distinguish active from inactive compounds for both ERα and ERβ, and (ii) build a web server for

138 discriminating compounds for estrogen receptor α and β with selectivity.

139

140 ## Materials & Methods

141 ### Data compilation and curation

142 Two data sets of inhibitors against ERα and ERβ were compiled from the ChEMBL database,

143 version 25 (Gaulton et al., 2017), which consisted of an initial set of 11,618 compounds for ERα

144 and 7,810 compounds for ERβ. The IC_{50} was selected as the bioactivity unit for further

145 investigation and after the data curation process, this led to a final data set of 1,593 and 1,281

146 compounds for ERα and ERβ, respectively. As this study sets out to develop a classification model

147 for both ERα and β we therefore, defined thresholds of <1 and >10 μM (corresponding to pIC_{50}

148 values of 6 and 5, respectively) for distinguishing actives from inactives, respectively. Moreover,

149 the intermediate biological activity with IC_{50} values ranging between 1 and 10 μM were not

150 selected for this study. A final set of non-redundant and curated compounds consisting of 1,194

151 and 997 inhibitors were obtained for ERα and ERβ, respectively.

152

153 ### Molecular descriptors

154

155 Fingerprint descriptors for compounds in the data sets were computed using the PaDEL-descriptor

156 software (Yap, 2011). SMILES notation was used for the calculation of molecular descriptors.

157 Structures were pre-processed so as to remove salt and standardize tautomers using the built-in

158 function of the PADEL-descriptor software. In general, molecular descriptors are important for

159 QSAR studies as they characterize molecular properties and chemical structure information in
160 quantitative or qualitative forms. As previously described in (Malik et al., 2020) 12 molecular
161 fingerprints belonging to 9 classes consisting of AtomPairs 2D, CDK fingerprinter, CDK extended,
162 CDK graph only, E-state, Klekota-Roth, MACCS, PubChem and Substructure were used for
163 describing the chemical structures. Furthermore, Substructure, Klekota-Roth fingerprint and 2D
164 atom pairs consisted of two versions: (1) 1 or 0 denotes the presence or absence of the descriptor
165 and the (2) representing the frequency value of the descriptor with a count version.
166

167 **Data filtering**
168 In order to remove inherent complexity and bias that may be introduced to the model building
169 process, constant and near constant variables were removed to select the fingerprint descriptor sets.
170 Particularly, near constants were identified using a standard deviation (SD) threshold of 0.1
171 whereby variables with SD values less than 0.1 were selected for further analysis.
172

173 **Data splitting**
174 The Kennard-Stone algorithm (Kennard & Stone, 1969) was applied for splitting the data into an
175 80/20 split where, the internal set comprised of 80% of the entire data set while the external set
176 consisted of the remaining 20%. The internal set is used as the training set and also subjected to 5-
177 fold CV. The external set is used as the testing set whereby the trained model will be applied to
178 this data set to make predictions and thus determine the model's robustness. For the internal set of
179 ER alpha (comprising 528 compounds), active compounds accounted for 53.59% while inactive
180 compounds are comprised of 46.41%. Similarly, for the external set (comprising of 131
181 compounds), active and inactive compounds were split into sets of 53.44% and 46.56%,
182 respectively. For ERβ, the percentage of active and inactive in the internal data set (composed of
183 572 compounds) was 78.15% and 21.85%, respectively. Similarly, in the external data set
184 (composed of 142 compounds), the percentage for active and inactive was 78.12% and 21.88%,
185 respectively.
186

187 **Statistical analysis**
188 As described in our previous work (Malik et. al., 2020), trends in individual descriptors of active
189 and inactive compounds were determined through 6 common descriptive statistical parameters,

190 encompassing the minimum (Min), first quartile (Q1), median, mean, third quartile (Q3) and

191 maximum (Max) parameters. The outcome was visualized in the form of a box plot using ggplot2,

192 a package of the R program. In addition, the Mann-Whitney U test (also known as the Wilcoxon

193 Rank Sum test) was conducted to determine the statistical significance in terms of the p-value.

194

195 **Multivariate analysis**

196

197 Random forest (RF) is used for the building of classification models owing to its robust model

198 performance and interpretability. RF was successfully used in our recent work for modeling the

199 bioactivity of hepatitis C virus inhibitors (Malik et. al., 2020). In essence, RF is an ensemble

200 classifier that employs an *N* number of decision trees (specified by the *ntree* parameter) to learn

201 the inherent patterns from the input data. In this study, a five-fold cross-validation (5-fold CV)

202 procedure was applied for tuning the *ntree* parameter (100, 1000, 100) and the *mtry* parameter (5,

203 30, 5) via the use of the tuneRF function from the *randomForest* package (Liaw & Wiener, 2002).

204 In order to provide a better understanding of the biochemical activity of the inhibitors, feature

205 selection was estimated using the built-in importance estimator of the RF model. The mean

206 decrease of the Gini index (MDGI) was utilized to estimate the important descriptors (Weidlich &

207 Filippov, 2016). Descriptors affording the largest value of MDGI represents the most important

208 features as that descriptor contributes most significantly to the model performance.

209

210 **Model validation**

211 Parameters commonly used for evaluating the model performance of binary classification

212 problems are typically based on true positives (TP), true negatives (TN), false positives (FP) and

213 false negatives (FN). Particularly, the fitness of the model was assessed using various statistical

214 parameters including the overall prediction accuracy (Ac), sensitivity (Sn), specificity (Sp) and

215 Matthews correlation coefficient (MCC) (Song & Tang, 2004).

216

217 $$Ac = TP + TN/(TP + TN + FP + FN) \times 100 \qquad (1)$$

218 $$Sn = TP/(TP + FN) \times 100 \qquad (2)$$

219 $$Sp = TN/(TN + FP) \times 100 \qquad (3)$$

220

221 $$MCC = (TP \times TN) - (FP \times FN)/\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)} \qquad (4)$$

222

223 where TP, TN, FP and FN represent the instances of true positives, true negatives, false positives

224 and false negatives, respectively.

225

226 **Applicability domain analysis**

227 The main purpose of the applicability domain (AD) is to estimate the boundaries within which the

228 model can make reliable and accurate predictions for compounds on the basis of similarity with

229 the compounds on which the model was constructed. The compounds that satisfy the scope of the

230 model are within the AD while the rest are outside the AD. In this study, we used the principal

231 component analysis (PCA) bounding box to assess the AD of compounds from the training

232 (internal) and testing (external) sets.

233

234 **Web server development**

235 The predictive model was exported as an RDS file (i.e., the model.RDS file) and subsequently

236 deployed as a web server. Particularly, the web server is coded in R via the use of the Shiny R

237 package (i.e., a web framework for the R environment). Technically, the web server is comprised

238 of two major components: (1) user interface and (2) server, which are saved as ui.R and server.R,

239 respectively. The ui.R file accepts input values (i.e. the SMILES notation of query compounds)

240 and transfers this information to the server.R file where the SMILES notation is submitted to the

241 PaDEL-Descriptor software. After descriptor calculation, the computed descriptors are used as

242 input to the predictive model (i.e., the model is exported as model.RDS file) which will then

243 classify query compounds as either being active or inactive (i.e., the bioactivity class label). Such

244 predicted class labels are then printed out onto the web server whereby users can also download

245 the predicted results as a CSV file.

246

247 **Reproducible research**

248 The data and code used in this study are publicly available on GitHub at

249 https://github.com/chaninlab/ERpred/.

250
251

Results and Discussion

A schematic summary of the workflow employed in this study is shown in Figure 2. Briefly, we start out by performing a chemical space analysis followed by QSAR model building and an in-depth feature analysis and finally deployed the best predictive model as a public web server.

Chemical space analysis

Chemical space analysis is employed to explore the characteristic differences between the active and inactive compounds. The general chemical space was first visualized as a function of the molecular weight (MW) versus the Ghose-Crippen-Viswanadhan octanol-water partition coefficient (ALogP). In addition, the active and inactive compounds were further compared using the Lipinski's rule-of-five (Ro5) descriptors. Briefly, the Ro5 describes the drug likeness of compounds on the basis of their molecular properties namely molecular weight (< 500), octanol–water partition coefficient (ALogP < 5), the number of hydrogen bond acceptors(< 10) and the number of hydrogen bond donors (< 5) (Lipinski et al., 2001). Visualization of the MW chemical space as a function of ALogP is shown in Figure 3. As can be observed for both ERα and ERβ, most of the compounds are clustered within the MW range of 200-500 Da with an ALogP in the range of 1 and 6. In addition, Figure 4 shows the distribution of active and inactive compounds according to the Ro5 descriptors. It is observed that both ER subtypes contain compounds following the Ro5 criteria such as, MW of less than 500 Da, ALogP value of less than 5 and nHBDon and nHBAcc values of less than 10. It can also be seen that for ERα, some of the active compounds have an ALogP of greater than 5, but the number is very minimal. Furthermore, the results from statistical analysis displays a significant difference between the active and inactive compounds using the Mann-Whitney U test. Most of the active compounds (422.68 ± 91.52) were larger (i.e., higher MW) than the inactive compounds (350.35 ± 79.82), which was observed from the mean values of box plots. Similarly, the ALogP values of the active compounds (4.36 ± 1.37) were greater than the inactive compounds (3.17 ± 1.53). However, it was observed that both active and inactive compounds had similar nHBDon values while the active compounds had nHBAcc values that were lower than the inactive compounds. On the other hand, for ERβ, the MW between the active (356.94 ± 92.43) and inactive compounds (351.69 ± 94.80) was not statistically significant as determined using the Mann-Whitney U test. Nonetheless, the ALogP was very statistically significant with the active group (3.82 ± 1.6) displaying higher values than the inactive

283   group (2.91 ± 1.5). Similar to the ERα subtype, the nHBDon values of both the active and inactive

284   groups were on par while the nHBAcc for the active compounds was seen to be a lot lower than

285   the inactive compounds.

286

287   Moreover, PubChem fingerprints were utilized for estimating the AD of the CSAR model

288   developed herein which were further used as input values for PCA analysis. The resulting PCA

289   scores plot can be seen in Figure 5. The data set for ERα and ERβ comprised of 1,194 and 997

290   compounds, respectively were further divided into internal (80%) and external sets (20%) using

291   the Kennard-Stone algorithm. Of note, the training set is composed of the internal set which is

292   utilized to build the model and thus make predictions on the external set. In addition, both the

293   internal and the external sets are also subjected to 5-fold CV. Furthermore, the chemical space

294   distribution as observed from Figure 5 shows that the external set (i.e., the testing set, represented

295   using blue dots) lies within the boundaries of the internal set (i.e., the training set, represented

296   using red dots). Thus, the AD is well defined for the CSAR model developed herein, as showns

297   through these results.

298

299   In order to develop a deeper understanding of the ER chemical space, the active and inactive

300   datasets for both ERα and ERβ were sorted according to their pIC$_{50}$ values. The top 10 active and

301   bottom 10 inactive compounds were taken from each set and applied to the Scaffold Hunter

302   software for further analysis (Schäfer et al., 2017). Particularly, the major scaffolds identified were

303   then observed in further detail in terms of the number of nodes in each parent scaffold lineages

304   and were also subjected to extensive analysis as will be discussed in paragraphs hereafter. Figure

305   6 shows the schematic process of the scaffold analysis.

306

307   As mentioned above, further in-depth exploration of ERα and ERβ inhibitors led to the

308   identification of top actives and inactives as shown in Supplementary Figures S1 and S2,

309   respectively. The top active compounds of ERα had a bioactivity range of 8.398-8.698 pIC$_{50}$ while

310   the top active compounds of ERβ had a bioactivity range of 8.045-8.522 pIC$_{50}$. Likewise, the

311   bioactivity range for the top inactive compounds pertaining to the ERα and ERβ groups were

312   observed have pIC$_{50}$ in the range of 3.008-3.252 and 3.000-3.587 pIC$_{50}$, respectively.

313

314   In addition, a rigorous analysis of the chemical space was conducted by investigating the
315   underlying scaffold structures as presented in actives and inactives using the Scaffold Hunter
316   software (Schäfer et al., 2017). Particularly, chemical structure clouds were created for actives and
317   inactives of ERα and ERβ as shown in Figure 7. Analysis of ERα actives led to the identification
318   of 2 major scaffolds namely 2,3-dihydro-1,4-benzoxathiine (with a frequency of 74) and
319   diaryltetrahydronaphthalene (with a frequency of 27). On the other hand, ERα inactives consisted
320   of 5 major scaffolds namely 4-hydroxypyrimidine (with a frequency of 26), 3-phenyl-4-
321   (3H)quinazolinone (having 4-hydroxypyrimidine as the parent with a frequency of 24), chromane
322   (with a frequency of 2), pyran-4-one (with a frequency of 8) and leucoline (with a frequency of
323   17). Similarly, analysis of ERβ actives led to the identification of 5 top scaffolds whereby 1 is in
324   common with ERα actives (i.e, lasofoxifene having a frequency of 3). The other 4 scaffolds
325   identified included 2-phenyl-1H-inden-1-one (with a frequency of 9), 2-phenylnaphthalene (with
326   a frequency of 59), 2-phenylbenzofuran (with a frequency of 21), and 1,2,9,9a-tetrahydrofluoren-
327   3-one (with a frequency of 25). Furthermore, 3 major scaffolds were observed for ERβ inactives,
328   which comprises of coumarone, pyrazolopyrimidine combined with a furan and 1,4-dioxa-8-
329   azaspiro[4.5]dec-8-yl(2-furyl)methanone. The importance of these above-mentioned scaffolds
330   will be discussed in the Structural interpretation section below.

331

332   **QSAR modeling**
333   This study follows the Organisation for Economic Co-operation and Development (OECD)
334   (OECD, 2014) guidelines for the development of robust QSAR models. These guidelines are
335   applied in all our work as previously mentioned (Malik et. al., 2020) and comprises of the
336   following main points: (i) the data set has a defined endpoint, (ii) uses an unambiguous learning
337   algorithm, (iii) the applicability domain of the QSAR model is well defined, (iv) appropriate
338   measures of goodness-of-fit, robustness and predictivity and (v) mechanistic interpretation of the
339   QSAR model. Following these aforementioned guidelines to develop interpretable QSAR models,
340   this study makes use of molecular fingerprints that are interpretable which are computed using the
341   PaDEL-Descriptor software. As shown in Table 1, 3 out of the 12 fingerprints (i.e., PubChem,
342   Substructure and Klekota-Roth) are readily interpretable. In addition, Table 2 provides details on
343   the model performances for all 12 fingerprints.

344

345   In this study, we developed a QSAR model based on the random forest algorithm in order to

346   differentiate the active and inactive inhibitors for ERα and ERβ subtypes. Table 2 shows the results

347   from the RF model with 12 different types of fingerprints over an internal validation test, CV and

348   an external validation test. The best averaged values were observed as Ac of 94.65% and 92.25%

349   and MCC of 0.89 and 0.76 for ERα and ERβ, respectively which was achieved for the PubChem

350   fingerprint descriptors as evaluated by CV. Concurrently, Klekota-Roth and Substructure

351   descriptors also performed well harbouring the second and third highest averaged values for Ac

352   and MCC in which models built using Klekota-Roth fingerprints afforded Ac and MCC values of

353   90.83% and 0.81, respectively for ERα and Ac and MCC values of 94.36% and 0.82, respectively

354   for ERβ. Similarly, models built using the Substructure fingerprints afforded Ac and MCC values

355   of 93.89% and 0.87, respectively for ERα and Ac and MCC values of 94.36% and 0.82,

356   respectively for ERβ. Although the Ac and MCC values of models built using the PubChem

357   fingerprints for ERβ were not superior to the models built using fingerprints of Klekota-Roth and

358   Substructure, they are quite comparable. In addition, taking into account the Ac and MCC values

359   as well as the overall external and CV, for both ERα and ERβ and the interpretability of the

360   features, we determined that the PubChem fingerprints were the ideal choice for interpretation of

361   the model.

362

363   **Mechanistic interpretation of feature importance**

364   In order to provide a better understanding of the mechanistic details governing ERα and β subtypes,

365   an analysis of the feature importance on selected informative descriptors was conducted. Owing

366   to the inbuilt ability of feature importance estimation of the RF model and its great prediction

367   performance, this study utilized it for analysis. Generally, two measures are used to rank the

368   important features, namely the mean decrease of the Gini index and the mean decrease of the

369   accuracy. As reported by Calle and Urrea (Calle & Urrea, 2011) that the Gini index had more

370   robust results compared to those from the accuracy, we utilized the mean decrease of the Gini

371   index to rank the importance of the PubChem feature descriptors. The 20 top-ranked PubChem

372   descriptors deduced from the Gini index as derived from the RF model can be found in Figure 8

373   and the contributions of their substructure towards the overall functioning of compounds as shown

374   in Tables 3 and 4, will be discussed in the following section.

375

376 For ERα, the Gini index pertaining to the top 20 ranked features are shown in Figure 8; and

377 described in Table 3, which consisted of descriptors pertaining to the following classes: 7 aromatic

378 (2 of which contain sulfur), 6 nitrogen containing features (consisting of amine and amide), 2 non-

379 aromatic sulfur containing compound and 5 aliphatic hydrocarbons or atom counts. In addition,

380 the Gini index pertaining to the top 20 ranked features for ERβ are shown in Figure 8 and described

381 in Table 4, which consisted of descriptors pertaining to the following classes: 10 nitrogen

382 containing features (consisting of amine, amide and atom count), 3 aromatic, 3 alcohol, 4 aliphatic

383 hydrocarbons or atom counts.

384

385 **Aromatic fingerprints**

386 For ERα, the maximum number of PubChem fingerprints as obtained from the Gini index, with 7

387 out of 20 top-ranked features (i.e., PubChemFP199, PubChemFP193, PubChemFP714,

388 PubChemFP777, PubChemFP259, PubChem804 and PubChemFP741) was seen to pertain to the

389 aromatic group. On the other hand, for ERβ, 3 out of the 20 top-ranked Gini features belonged to

390 the aromatic group (i.e., PubChemFP777, PubChemFP714 and PubChemFP259). Surprisingly, all

391 of the 3 aromatic Gini features for ERβ were also present for ERα. Therefore, we can infer that

392 these 3 aromatic features are important to the functioning of the compounds. Diving deeper into

393 the substructure description of these PubChem features, it can be seen that the first, second and

394 ninth ranked features, (i.e., PubChem199, PubChem193 and PubChem259) correspond to aromatic

395 rings of size $\geq$ 3 or 4. This point is in accordance to the fact that the natural agonist of ER (i.e.

396 estradiol or E2) has only 1 aromatic ring while the most common antagonists of ER (i.e. tamoxifen

397 and fulvestrant) have 3 and 1 aromatic rings, respectively (Bafna et al., 2020). In addition, it is the

398 phenyl group in E2 that forms hydrogen bonds with Glu353 and Arg394 in the active site.

399 Similarly, tamoxifen with its triphenylethylene core forms the same hydrogen bonds as E2 (Bafna

400 et al., 2020). Thus, the phenyl moiety of compounds are important for ER inhibition. In addition,

401 the third and seventh ranked features i.e., PubChem714 and PubChem777, correspond to 4-

402 methylphenol which is an organic compound used as a precursor or intermediate for the

403 manufacture of other chemicals. Furthermore, 4-methylphenol is also vital in the production of

404 butylated hydroxytoluene (BHT) undergoing coupling to give an extensive family of diphenol

405 antioxidants. These antioxidants are valued because they are relatively low in toxicity (Fiege,

406 2000). Although BHT has also been postulated as an antiviral drug, it has not yet been approved

407   by any drug regulatory agency for use as an antiviral (Pirtle, Sacks & Nachman, 1986; Lanigan &

408   Yamarik, 2002).

409

410   Furthermore, the tenth and fifteenth ranked features (i.e. PubChem804 and PubChem741)

411   correspond to 3-sulfonyl phenol which according to the SMILES from its substructure description,

412   seems to fit as a part of 4,4'-sulfonyldiphenol (Bisphenol S) (National Center for Biotechnology

413   Information, 2020). As described in the report, Bisphenol S is an organic compound that has many

414   functions, one of them being to act as an endocrine disruptor and it could thus modulate hormone

415   receptors such as ERα and ERβ (Rochester & Bolden, 2015). For example, Viñas and Watson

416   (Viñas & Watson, 2013) studied the nongenomic effects of Bisphenol S since it is an analogue of

417   Bisphenol A, a well-known endocrine disruptor that imperfectly mimics the effects of physiologic

418   estrogens via membrane-bound estrogen receptors. The authors concluded that Bisphenol S

419   disrupts E2-induced cell signaling, leading to altered cell proliferation and cell death. Hence, the

420   presence of such features in the Gini index top 20 is valid.

421

422   **Nitrogen-containing fingerprints**

423   According to the Gini index for ERβ, 10 out of the top 20 features pertain to the nitrogen containing

424   class which includes PubChemFP392, PubChemFP451, PubChemFP439, PubChemFP393,

425   PubChemFP299, PubChemFP450, PubChemFP645, PubChemFP15, PubChemFP375 and

426   PubChemFP646. On the other hand, features containing nitrogen as obtained from the Gini index

427   for ERα constituted of 6 features, namely, PubChemFP345, PubChemFP540, PubChemFP365,

428   PubChemFP453, PubChemFP391 and PubChemFP418. Taken together, features with nitrogen

429   (i.e., amines and amides) constituted the maximum number of features, spanning both ER

430   subtypes. Interestingly, there were no overlapping features found between the two groups for this

431   category. Furthermore, the substitution of the CH group with a N atom in compounds containing

432   aromatic and heteroaromatic ring systems, is a common bioisosteric transformation conducted to

433   mimic the binding of natural ligands while exerting antagonistic effects (Kumar et al., 2011). Of

434   note, 6 out of the 10 nitrogen containing features for ERβ (Table 4) belonged to the top 10 of Gini

435   index, emphasizing their importance. In addition, most of the aforementioned features pertain to

436   methanamine, *N*-methylmethanamine, ethylamine, *N,N*-dimethylmethanamine etc. which are all

437   precursors of many significant chemical compounds such as Tamoxifen, 4-hydroxy tamoxifen,

438   Raloxifene and many of their derived analogues. A review by Sharma et al (Sharma, Kumar &

439   Narasimhan, 2018) highlights the substructures of ERα antagonists and their analogues which

440   were analyzed *in silico* using molecular docking. Through this review, the authors emphasized the

441   need for selective estrogen receptor antagonists for the treatment of breast cancer.

442

443   **Aliphatic hydrocarbons**

444   From the analysis of the Gini index, for ERα and ERβ, respectively, 2 and 4 features out of the top

445   20 ranked features belonged to this group. PubChem697 and PubChem696 corresponding to 2-

446   methlyheptane (which is also isomeric to octane) and octane, respectively were observed to be a

447   common feature for both ER subtypes (Tables 3 and 4). In addition, Shoda et al (Shoda et al.,

448   2015) observed that long alkyl side chains of tamoxifen derivatives acted as inducers of ER

449   degradation. Furthermore, the same group designed a series of diphenylalkane derivatives bearing

450   several long alkyl chains on the hydroxyl group and evaluated their biological properties such as

451   ER degradation, binding affinity, transcriptional activity and anti-proliferation activity (Shoda et

452   al., 2015; Misawa et al., 2017). Among all the compounds, one emerged as a novel ERα

453   downregulator with a binding affinity of $IC_{50} = 4.9$ nM and ERα antagonistic activity of $IC_{50} = 45$

454   nM. Moreover, upon conducting computational docking analysis of the novel compound, the

455   authors observed the interactions between the hydrogen atom on the amino group of the compound

456   and the carboxylic acid of Glu351 of ERα which leads to the binding of the long alkyl chain to the

457   hydrophobic groove of ERα. Thus, the amino group and the optimal length of the long alkyl chain

458   in the diphenylheptane skeleton are considered important for ERα downregulation (Nanjyo et al.,

459   2019).

460

461   **Structural interpretation**

462   As previously mentioned, the LBD of ERα and ERβ show a consensus of around 55% and are

463   composed of 12 α helices. Upon interaction with its natural ligand (i.e. E2), both ER subtypes form

464   hydrogen bonds with residues (ERα/β numbering) Glu353/305, His524/475, Arg394/346 and a

465   water bridge for Arg394/346 which connects to the A-ring of E2 (Salentin et al., 2015). In addition,

466   both ERs also form perpendicular pi-stacking with the phenol rings of residue Phe404/356 and E2.

467   Similarly, hydrophobic interactions were also observed with the interactions of both ER subtypes

468   to E2 (consisting of residues M343, Leu346, Leu349, Ala350, Leu387 and Leu391 for ERα and

469   Leu298, Leu301, Ala302, M336, Leu339, Leu343 and Leu476 for ERβ) however, ERβ possesses

470   an additional interaction with residue Leu476 as compared to ERα (Salentin et al., 2015). Figure

471   9 shows these aforementioned interactions, although some interactions are not shown for

472   simplicity.

473

474   Generally, a good ER antagonist should possess two OH groups that are linked by a lipophilic

475   central scaffold, which places them at a distance of about 11 Å. At least one of these hydroxyls

476   should be a phenol or a phenol-bioisostere (i.e. replacement of the phenol group with another group

477   that can act as a similar hydrogen bond donor) (Sessler et al., 2017). One of these OH groups form

478   a strong hydrogen bond with residues (ERα/β numbering) Glu353/305, Arg394/346 and a water

479   molecule and the other OH group, instead, mimics the OH group of estradiol and forms an

480   additional hydrogen bond with residue His524/475 in the estradiol D-ring pocket. At the other end

481   of the binding cavity, the D-ring makes non-polar contacts with Ile424/376, Gly521/472 and

482   Leu525/476 (Brzozowski et al., 1997). Beyond these key, energetically important interactions,

483   which appear to be comparable between the two ERs, subtype-selectivity needs to arise from the

484   different shape and hydrophobicity of the central scaffold. This has previously been demonstrated

485   by Pike et al. (Pike et al., 1999), where the authors study the interactions of ERβ-Genistein (a ERβ-

486   specific partial agonist) complex and observed that the H12 helix does not adopt the distinctive

487   'agonist' conformation but instead, lies in the 'antagonistic' position. In addition, it was further

488   observed that two amino acid changes within the binding cavity (i.e., residues Leu384/Met336,

489   and Met421/Ile373) may be responsible for specificity towards ERβ as well as differences in the

490   propensity of the two ERs to form pockets of different sizes. In any case, the selective activation

491   of ERα or ERβ may depend not only on a selective receptor binding affinity, but also on selective

492   activation of each receptor subtype (Leitman et al., 2010).

493

494   Several studies have reported a novel class of Selective Estrogen Receptor Modulators (SERMs)

495   based on dihydrobenzoxathiine (i.e., the top scaffold for ERα actives) as shown in Figure 7 (Kim

496   et al., 2004; Chen et al., 2004; Viglianisi & Menichetti, 2010). A SAR evaluation of several analogs

497   was carried out for identifying the best structural features for SERM both *in vitro* and *in vivo*. It

498   appears that the sulfur atom of the dihydrobenzoxathiine scaffold interacts with discriminative

499   residues (Leu384 for ERα and Met336 for ERβ) in the binding pocket of the two receptor isoforms

500   and thus plays a crucial role in maintaining the subtype-selectivity. This particular scaffold was
501   observed in 5 of the top 10 active compounds (i.e., compounds **4**, **7**, **8**, **9** and **10** of Supplementary
502   Figure S1) and therefore, its importance in preserving high activity towards ERα is evident.
503   Another study revealed the diastereomerism of dihydrobenzoxathiine through molecular modeling
504   (Zhuang et al., 2011) and discovered that the full antagonistic activity was achieved through
505   hydrogen bonding with Glu353 and His524 of ERα LBD while van der Waals interactions were
506   most predominant in the binding. In addition, lasofoxifene (derived from
507   diaryltetrahydronaphthalene) was seen to be another major scaffold found in both ERα and ERβ
508   active groups. Lasofoxifene selectively binds to both ERs with high affinity that is found to be
509   similar to that of estradiol as well as other reported SERMs (i.e., raloxifene and tamoxifen)
510   (Gennari et al., 2006). Furthermore, the crystal structure of lasofoxifene bound to ERα revealed
511   features of ERα/SERM recognition whereby the C-terminal AF-2 helix is displaced at the LXXLL
512   motif of coactivator proteins (occupying the space normally filled by residue Leu 540) as well as
513   modulating the conformation of helix 11 residues (His 524, Leu 525) (Vajdos et al., 2007).

514

515   Additionally, the 2-phenylnaphthalene scaffold was seen to have a high frequency number for ERβ
516   (Supplementary Figure S2). Genistein is a well known ERβ selective inhibitor that is derived from
517   the 2-phenylnaphthalene scaffold. Furthermore, genistein is a key compound in the further
518   identification of ERβ specific inhibitors (Mewshaw et al., 2005). The authors introduced
519   substitutions at the appropriate positions of the phenylnaphthalene scaffold allowing a single
520   orientation to predominate, which accounted for higher ERβ selectivity. Furthermore, Wilkening
521   et al. (2006) noted that by creating analogues of tetrahydrofluorenone through substitution, potent
522   ERβ subtype selective ligands were formed. It should be noted that tetrahydrofluorenone
523   accounted for 3 of the top 10 from ERβ actives (i.e., compounds **23**, **25** and **26**) as shown in
524   Supplementary Figure S2. Furthermore, the authors had also reported several analogs possessing
525   ERβ binding affinities that is comparable to that of 17β-estradiol but with greater than 75-fold
526   selectivity over that of ERα. Therefore, this analysis could help guide the process of novel ER
527   inhibitor development that are subtype-selective.

528

529   **Model deployment as a web server**
530   In order to make the prediction model presented herein a practical tool that can be widely used by

531 the scientific community, we have constructed a web server called the ERpred using the model as

532 described in previous sections. Briefly, the SMILES notation of the chemical compound of interest

533 is used as the input which is fed into the ERpred web server. The server performs descriptor

534 calculation using the PaDEL-Descriptor software and outputs predictions in the form of the

535 bioactivity class label based upon the constructed random forest model. In addition, other web

536 servers established for bioactivity prediction do so for just one class however, the ERpred web

537 server is unique since it can predict the bioactivity for 2 proteins (herein ERα and ERβ),

538 simultaneously. The Shiny package under the R programming environment was used to establish

539 the web interface which has been made freely accessible at http://codes.bio/erpred/. Figure 10

540 provides screenshots of the ERpred web server where the web server prior to input submission

541 (i.e., panel A) and after prediction (i.e., panel B) are clearly depicted.

542

543 Briefly, a step-by-step guide on using the web server is given below:

544 ● Step 1. The following URL, http://codes.bio/ERpred/ should be entered into the web
545   browser.

546 ● Step 2. SMILES notation for common compounds of interest can be obtained from public
547   databases such as ChEMBL, PubChem or ChemSpider whereas custom molecules can also
548   be drawn into ChemDraw or ChemAxon MarvinSketch so as to generate the SMILES
549   notation.

550 ● Step 3. Input molecules as SMILES notation should be entered into the Input box or
551   uploaded as a file containing the SMILES notation by clicking on the "Choose file" button.

552 ● Step 4. Prediction process can be initiated upon clicking the "Submit" button.

553 ● Step 5. Prediction results are automatically displayed in the grey box found below the
554   "Status/Output" heading. Typically, it takes a few seconds for the server to process the
555   task. Users can also download the prediction results as a CSV file by pressing on the
556   "Download CSV button".

557

558 ## Conclusions

559 Breast cancer is the most frequently detected cancer among women with over 2 million new cases

560 and an estimated 627,000 deaths (15% of all cancer deaths in women) in 2018 (World Health

561 Organization, 2018). However, current inhibitors and hormone therapy is problematic due to the

562   development of resistance and an increased risk of breast and endometrial cancers, and

563   thromboembolism. Thus, this study had qualitatively and quantitatively addressed these issues by

564   building a QSAR model capable of distinguishing active compounds from inactive compounds for

565   ERα and ERβ. In addition, to allow researchers from all backgrounds easy access to our prediction

566   model, we built a web server for discriminating compounds for ERα and ERβ with selectivity. ER

567   activity prediction was evaluated via machine learning algorithms and several classes of

568   fingerprint descriptors. The results obtained indicated that the RF algorithm coupled with the

569   PubChem fingerprints allowed for the most interpretable descriptors along with the best

570   performing model. The feature analysis of important substructure contributions as obtained from

571   the Gini index revealed that aromaticity, amine groups and aliphatic hydrocarbons were important

572   for the active compounds. Moreover, in-depth scaffold analysis of the top active compounds

573   revealed that the binding specificity for ERα and ERβ involve different scaffolds. Since not many

574   studies have focused on ERβ, this protein is worth further explorations. Thus, the knowledge

575   gained from this study serves as general guidelines for the data driven design of potentially active

576   and selective estrogen receptor antagonists.

577

Acknowledgements

579   This work was supported by the New Researcher Grant (A31/2561) from Mahidol University as

580   well as the TRF Research Career Development Grant (No. RSA6280075) from the Thailand

581   Research Fund, the Office of Higher Education Commission and Mahidol University.

582

Reference

584   Abdulkareem IH., Zurmi IB. 2012. Review of hormonal treatment of breast cancer. *Nigerian journal of*
585   *clinical practice* 15:9–14.
586   Althuis MD., Fergenbaum JH., Garcia-Closas M., Brinton LA., Madigan MP., Sherman ME. 2004.
587   Etiology of hormone receptor-defined breast cancer: a systematic review of the literature. *Cancer*
588   *Epidemiology, Biomarkers & Prevention* 13:1558–1568.
589   Arao Y., Hamilton KJ., Coons LA., Korach KS. 2013. Estrogen receptor α L543A,L544A mutation changes
590   antagonists to agonists, correlating with the ligand binding domain dimerization associated with
591   DNA binding activity. *The Journal of Biological Chemistry* 288:21105–21116.
592   Bafna D., Ban F., Rennie PS., Singh K., Cherkasov A. 2020. Computer-Aided Ligand Discovery for
593   Estrogen Receptor Alpha. *International Journal of Molecular Sciences* 21.
594   Ballester PJ. 2019. Machine learning for molecular modelling in drug design. *Biomolecules* 9.
595   Breiman L. 2001. Random Forests. *Springer Science and Business Media LLC*.

596   Breiman L., Friedman JH., Olshen RA., Stone CJ. 1984. Classification and regression trees. *Biometrics*
597   40:874.
598   Brzozowski AM., Pike AC., Dauter Z., Hubbard RE., Bonn T., Engström O., Ohman L., Greene GL.,
599   Gustafsson JA., Carlquist M. 1997. Molecular basis of agonism and antagonism in the oestrogen
600   receptor. *Nature* 389:753–758.
601   Calle ML., Urrea V. 2011. Letter to the editor: Stability of Random Forest importance measures. *Briefings*
602   *in Bioinformatics* 12:86–89.
603   Chen HY., Kim S., Wu JY., Birzin ET., Chan W., Yang YT., Dahllund J., DiNinno F., Rohrer SP., Schaeffer
604   JM., Hammond ML. 2004. Estrogen receptor ligands. Part 3: The SAR of dihydrobenzoxathiin
605   SERMs. *Bioorganic & Medicinal Chemistry Letters* 14:2551–2554.
606   Cherkasov A., Muratov EN., Fourches D., Varnek A., Baskin II., Cronin M., Dearden J., Gramatica P.,
607   Martin YC., Todeschini R., Consonni V., Kuz'min VE., Cramer R., Benigni R., Yang C., Rathman
608   J., Terfloth L., Gasteiger J., Richard A., Tropsha A. 2014. QSAR modeling: where have you been?
609   Where are you going to? *Journal of Medicinal Chemistry* 57:4977–5010.
610   Cook KL., Shajahan AN., Clarke R. 2011. Autophagy and endocrine resistance in breast cancer. *Expert*
611   *Review of Anticancer Therapy* 11:1283–1294.
612   Coriano CG., Liu F., Sievers CK., Liang M., Wang Y., Lim Y., Yu M., Xu W. 2018. A Computational-
613   Based Approach to Identify Estrogen Receptor α/β Heterodimer Selective Ligands. *Molecular*
614   *Pharmacology* 93:197–207.
615   Dahlman-Wright K., Cavailles V., Fuqua SA., Jordan VC., Katzenellenbogen JA., Korach KS., Maggi A.,
616   Muramatsu M., Parker MG., Gustafsson J-A. 2006. International Union of Pharmacology. LXIV.
617   Estrogen receptors. *Pharmacological Reviews* 58:773–781.
618   Fiege H. 2000. Cresols and Xylenols. In: Wiley-VCH Verlag GmbH & Co. KGaA ed. *Ullmann's*
619   *encyclopedia of industrial chemistry*. Weinheim, Germany: Wiley-VCH Verlag GmbH & Co.
620   KGaA,.
621   Foulkes WD., Smith IE., Reis-Filho JS. 2010. Triple-negative breast cancer. *The New England Journal of*
622   *Medicine* 363:1938–1948.
623   Gamucci T., Vaccaro A., Ciancola F., Pizzuti L., Sperduti I., Moscetti L., Longo F., Fabbri MA., Giampaolo
624   MA., Mentuccia L., Di Lauro L., Vici P. 2013. Recurrence risk in small, node-negative, early breast
625   cancer: a multicenter retrospective analysis. *Journal of Cancer Research and Clinical Oncology*
626   139:853–860.
627   Gaulton A., Hersey A., Nowotka M., Bento AP., Chambers J., Mendez D., Mutowo P., Atkinson F., Bellis
628   LJ., Cibrián-Uhalte E., Davies M., Dedman N., Karlsson A., Magariños MP., Overington JP.,
629   Papadatos G., Smit I., Leach AR. 2017. The ChEMBL database in 2017. *Nucleic Acids Research*
630   45:D945–D954.
631   Gennari L., Merlotti D., Martini G., Nuti R. 2006. Lasofoxifene: a third-generation selective estrogen
632   receptor modulator for the prevention and treatment of osteoporosis. *Expert Opinion on*
633   *Investigational Drugs* 15:1091–1103.
634   Han J., Lim W., You D., Jeong Y., Kim S., Lee JE., Shin TH., Lee G., Park S. 2019. Chemoresistance in
635   the Human Triple-Negative Breast Cancer Cell Line MDA-MB-231 Induced by Doxorubicin
636   Gradient Is Associated with Epigenetic Alterations in Histone Deacetylase. *Journal of Oncology*
637   2019:1345026.
638   Heldring N., Pike A., Andersson S., Matthews J., Cheng G., Hartman J., Tujague M., Ström A., Treuter E.,
639   Warner M., Gustafsson JA. 2007. Estrogen receptors: how do they signal and what are their targets.

640   *Physiological Reviews* 87:905–931.

641   Hiscox S., Davies EL., Barrett-Lee P. 2009. Aromatase inhibitors in breast cancer. *Maturitas* 63:275–279.

642   Ji X., Lu Y., Tian H., Meng X., Wei M., Cho WC. 2019. Chemoresistance mechanisms of breast cancer
643   and their countermeasures. *Biomedicine & Pharmacotherapy (Biomedecine & Pharmacotherapie)*
644   114:108800.

645   Jordan VC. 2001. Selective estrogen receptor modulation: a personal perspective. *Cancer Research*
646   61:5683–5687.

647   Kammerer M., Gutzwiller S., Stauffer D., Delhon I., Seltenmeyer Y., Fournier B. 2013. Estrogen Receptor
648   α (ERα) and Estrogen Related Receptor α (ERRα) are both transcriptional regulators of the Runx2-I
649   isoform. *Molecular and Cellular Endocrinology* 369:150–160.

650   Kennard RW., Stone LA. 1969. Computer aided design of experiments. *Technometrics : a journal of*
651   *statistics for the physical, chemical, and engineering sciences* 11:137–148.

652   Kerdivel G., Habauzit D., Pakdel F. 2013. Assessment and molecular actions of endocrine-disrupting
653   chemicals that interfere with estrogen receptor pathways. *International journal of endocrinology*
654   2013:501851.

655   Kim SI., Sohn J., Koo JS., Park SH., Park HS., Park BW. 2010. Molecular subtypes and tumor response to
656   neoadjuvant chemotherapy in patients with locally advanced breast cancer. *Oncology* 79:324–330.

657   Kim S., Wu JY., Birzin ET., Frisch K., Chan W., Pai L-Y., Yang YT., Mosley RT., Fitzgerald PMD.,
658   Sharma N. et al. 2004. Estrogen receptor ligands. II. Discovery of benzoxathiins as potent, selective
659   estrogen receptor alpha modulators. *Journal of Medicinal Chemistry* 47:2171–2175.

660   Kumar R., Zakharov MN., Khan SH., Miki R., Jang H., Toraldo G., Singh R., Bhasin S., Jasuja R. 2011.
661   The dynamic structure of the estrogen receptor. *Journal of amino acids* 2011:812540.

662   Lanigan RS., Yamarik TA. 2002. Final report on the safety assessment of BHT(1). *International journal of*
663   *toxicology* 21 Suppl 2:19–94.

664   Leitman DC., Paruthiyil S., Vivar OI., Saunier EF., Herber CB., Cohen I., Tagliaferri M., Speed TP. 2010.
665   Regulation of specific target genes and biological responses by estrogen receptor subtype agonists.
666   *Current Opinion in Pharmacology* 10:629–636.

667   Leung Y-K., Mak P., Hassan S., Ho S-M. 2006. Estrogen receptor (ER)-beta isoforms: a key to
668   understanding ER-beta signaling. *Proceedings of the National Academy of Sciences of the United*
669   *States of America* 103:13162–13167.

670   Liaw A., Wiener M. 2002. Classification and Regression by randomForest. *RNews* 2:18–22.

671   Lipinski CA., Lombardo F., Dominy BW., Feeney PJ. 2001. Experimental and computational approaches
672   to estimate solubility and permeability in drug discovery and development settings. *Advanced Drug*
673   *Delivery Reviews* 46:3–26.

674   Malik AA., Phanus-umporn P., Schaduangrat N., Shoombuatong W., Isarankura-Na-Ayudhya C.,
675   Nantasenamat C. 2020. HCVpred: A web server for predicting the bioactivity of hepatitis C virus
676   NS5B inhibitors. *Journal of Computational Chemistry* 48:1820–1834.

677   Manas ES., Xu ZB., Unwalla RJ., Somers WS. 2004. Understanding the selectivity of genistein for human
678   estrogen receptor-beta using X-ray crystallography and computational methods. *Structure*
679   12:2197–2207.

680   Mewshaw RE., Edsall RJ., Yang C., Manas ES., Xu ZB., Henderson RA., Keith JC., Harris HA. 2005.
681   ERbeta ligands. 3. Exploiting two binding orientations of the 2-phenylnaphthalene scaffold to
682   achieve ERbeta selectivity. *Journal of Medicinal Chemistry* 48:3953–3979.

683   Misawa T., Fujisato T., Kanda Y., Ohoka N., Shoda T., Yorioka M., Makishima M., Sekino Y., Naito M.,

684   Demizu Y., Kurihara M. 2017. Design and synthesis of novel selective estrogen receptor
685   degradation inducers based on the diphenylheptane skeleton. *MedChemComm* 8:239–246.
686 Nanjyo S., Ohgane K., Yoshioka H., Makishima M., Hashimoto Y., Noguchi-Yachide T. 2019. Structure-
687   activity relationship study of estrogen receptor down-regulators with a diphenylmethane skeleton.
688   *Bioorganic & Medicinal Chemistry* 27:1952–1961.
689 Nantasenamat C., Isarankura-Na-Ayudhya C., Prachayasittikul V. 2010. Advances in computational
690   methods to predict the biological activity of compounds. *Expert opinion on drug discovery* 5:633–
691   654.
692 National Center for Biotechnology Information. 2020. PubChem Compound Summary for CID 6626, 4,4'-
693   Sulfonyldiphenol.
694 OECD. 2014. *Guidance Document on the Validation of (Quantitative) Structure-Activity Relationship*
695   *[(Q)SAR] Models*. OECD.
696 Osborne CK., Wakeling A., Nicholson RI. 2004. Fulvestrant: an oestrogen receptor antagonist with a novel
697   mechanism of action. *British Journal of Cancer* 90 Suppl 1:S2-6.
698 Pang X., Fu W., Wang J., Kang D., Xu L., Zhao Y., Liu A-L., Du G-H. 2018. Identification of Estrogen
699   Receptor α Antagonists from Natural Products via In Vitro and In Silico Approaches. *Oxidative*
700   *medicine and cellular longevity* 2018:6040149.
701 Paterni I., Granchia C., Katzenellenbogen JA., Minutolo F. 2014. Estrogen receptors alpha (ERα) and beta
702   (ERβ): Subtype-selective ligands and clinical potential. *Steroids* 90: 13–29.
703 Pavlin M., Spinello A., Pennati M., Zaffaroni N., Gobbi S., Bisi A., Colombo G., Magistrato A. 2018. A
704   computational assay of estrogen receptor α antagonists reveals the key common structural traits of
705   drugs effectively fighting refractory breast cancers. *Scientific Reports* 8:649.
706 Pike AC., Brzozowski AM., Hubbard RE., Bonn T., Thorsell AG., Engström O., Ljunggren J., Gustafsson
707   JA., Carlquist M. 1999. Structure of the ligand-binding domain of oestrogen receptor beta in the
708   presence of a partial agonist and a full antagonist. *The EMBO Journal* 18:4608–4618.
709 Pirtle EC., Sacks JM., Nachman RJ. 1986. Antiviral effectiveness of butylated hydroxytoluene against
710   pseudorabies (Aujeszky's disease) virus in cell culture, mice, and swine. *American Journal of*
711   *Veterinary Research* 47:1892–1895.
712 Ramirez-Ardila DE., Helmijr JC., Look MP., Lurkin I., Ruigrok-Ritstier K., van Laere S., Dirix L., Sweep
713   FC., Span PN., Linn SC. et al. 2013. Hotspot mutations in PIK3CA associate with first-line
714   treatment outcome for aromatase inhibitors but not for tamoxifen. *Breast Cancer Research and*
715   *Treatment* 139:39–49.
716 Rochester JR., Bolden AL. 2015. Bisphenol S and F: A systematic review and comparison of the hormonal
717   activity of bisphenol A substitutes. *Environmental Health Perspectives* 123:643–650.
718 Salentin S., Schreiber S., Haupt VJ., Adasme MF., Schroeder M. 2015. PLIP: fully automated protein-
719   ligand interaction profiler. *Nucleic Acids Research* 43:W443-7.
720 Schäfer T., Kriege N., Humbeck L., Klein K., Koch O., Mutzel P. 2017. Scaffold Hunter: a comprehensive
721   visual analytics framework for drug discovery. *Journal of cheminformatics* 9:28.
722 Sessler CD., Rahm M., Becker S., Goldberg JM., Wang F., Lippard SJ. 2017. CF2H, a hydrogen bond
723   donor. *Journal of the American Chemical Society* 139:9325–9332.
724 Sharma D., Kumar S., Narasimhan B. 2018. Estrogen alpha receptor antagonists for the treatment of breast
725   cancer: a review. *Chemistry Central Journal* 12:107.
726 Shiau AK., Barstad D., Radek JT., Meyers MJ., Nettles KW., Katzenellenbogen BS., Katzenellenbogen
727   JA., Agard DA., Greene GL. 2002. Structural characterization of a subtype-selective ligand reveals

728   a novel mode of estrogen receptor antagonism. *Nature Structural Biology* 9:359–364.

729 Shoda T., Kato M., Harada R., Fujisato T., Okuhira K., Demizu Y., Inoue H., Naito M., Kurihara M. 2015.
730   Synthesis and evaluation of tamoxifen derivatives with a long alkyl side chain as selective estrogen
731   receptor down-regulators. *Bioorganic & Medicinal Chemistry* 23:3091–3096.

732 Song J., Tang H. 2004. Accurate classification of homodimeric vs other homooligomeric proteins using a
733   new measure of information discrepancy. *Journal of Chemical Information and Computer Sciences*
734   44:1324–1327.

735 Thrane S., Lykkesfeldt AE., Larsen MS., Sorensen BS., Yde CW. 2013. Estrogen receptor α is the major
736   driving factor for growth in tamoxifen-resistant breast cancer and supported by HER/ERK
737   signaling. *Breast Cancer Research and Treatment* 139:71–80.

738 Vajdos FF., Hoth LR., Geoghegan KF., Simons SP., LeMotte PK., Danley DE., Ammirati MJ., Pandit J.
739   2007. The 2.0 A crystal structure of the ERalpha ligand-binding domain complexed with
740   lasofoxifene. *Protein Science* 16:897–905.

741 Viglianisi C., Menichetti S. 2010. Dihydrobenzo[1,4]oxathiine: A multi-potent pharmacophoric
742   heterocyclic nucleus. *Current Medicinal Chemistry* 17:915–928.

743 Viñas R., Watson CS. 2013. Bisphenol S disrupts estradiol-induced nongenomic signaling in a rat pituitary
744   cell line: effects on cell functions. *Environmental Health Perspectives* 121:352–358.

745 Weidlich IE., Filippov IV. 2016. Using the gini coefficient to measure the chemical diversity of small-
746   molecule libraries. *Journal of Computational Chemistry* 37:2091–2097.

747 Weiser MJ., Foradori CD., Handa RJ. 2008. Estrogen receptor beta in the brain: from form to function.
748   *Brain Research Reviews* 57:309–320.

749 Welboren W-J., Sweep FCGJ., Span PN., Stunnenberg HG. 2009. Genomic actions of estrogen receptor
750   alpha: what are the targets and how are they regulated? *Endocrine-Related Cancer* 16:1073–1089.

751 Wilkening RR., Ratcliffe RW., Tynebor EC., Wildonger KJ., Fried AK., Hammond ML., Mosley RT.,
752   Fitzgerald PMD., Sharma N., McKeever BM. et al. 2006. The discovery of tetrahydrofluorenones
753   as a new class of estrogen receptor beta-subtype selective ligands. *Bioorganic & Medicinal*
754   *Chemistry Letters* 16:3489–3494.

755 Wójcikowski M., Siedlecki P., Ballester PJ. 2019. Building Machine-Learning Scoring Functions for
756   Structure-Based Prediction of Intermolecular Binding Affinity. *Methods in Molecular Biology*
757   2053:1–12.

758 World Health Organization. 2018. Breast cancer.

759 Yap CW. 2011. PaDEL-descriptor: an open source software to calculate molecular descriptors and
760   fingerprints. *Journal of Computational Chemistry* 32:1466–1474.

761 Zhuang S., Zhang J., Zhang F., Zhang Z., Wen Y., Liu W. 2011. Investigation of the diastereomerism of
762   dihydrobenzoxathiin SERMs for ER alpha by molecular modeling. *Bioorganic & Medicinal*
763   *Chemistry Letters* 21:7298–7305.

764

765

766 **Figure Legends**

767

768 **Figure 1. Protein structure and sequence alignment of ER subtypes.** ERα and ERβ are

769 displayed in blue and yellow colors, respectively. Secondary structure elements consisting of 12
770 helices and 2 strands are shown. Sequence alignment of the ligand binding domain of the two ER
771 subtypes were performed in Clustal Omega.

772

773 **Figure 2. Schematic representation of the methodological workflow of this study.**

774

775 **Figure 3. Plot of MW vs ALogP for compounds in the ERα and ERβ datasets.** The plot
776 allows simple visualization of the chemical space of inhibitors against ERα (A) and ERβ (B).
777 Active and inactive compounds are shown in salmon pink and teal colors, respectively.

778

779 **Figure 4. Box plot of Lipinski's rule-of-five descriptors.** The four rule-of-five descriptors are
780 shown for the ERα (A) and ERβ (B) datasets. Active and inactive compounds are shown in
781 salmon pink and teal colors, respectively.

782

783 **Figure 5. PCA scores plot for compounds in the ERα and ERβ datasets.** The scores plot
784 allows visualization of the distribution of compounds for internal (blue) and external (red) sets
785 that constitutes the assessment of the applicability domain for ERα (A) and ERβ (B) datasets.

786

787 **Figure 6. Schematic representation of the methodological workflow of obtaining the scaffolds
788 for ERα and ERβ active and inactive groups.** Top active and inactive compounds were
789 determined from their pIC_{50} values. Scaffold Hunter was used to create scaffold trees whereby top
790 scaffolds were determined for each bioactivity class (i.e., actives and inactives) for both ERα and
791 ERβ.

792

793 **Figure 7. Chemical structures cloud of ERα and ERβ actives and inactives.** Chemical structure
794 cloud of actives (top panels) and inactives (bottom panels) for ERα and ERβ inhibitors.
795 Particularly, active compounds were defined as compounds having IC_{50} in the range of 10-1000
796 nM whereby highly actives (> 10 nM) are represented by salmon pink color while weakly actives
797 (< 10 nM) are represented in grey. Inactive compounds were defined as compounds having IC_{50} in
798 the range of 10,000-1,000,000 nM whereby the teal color corresponds to the most inactive
799 compounds (towards the 1,000,000 nM scale) while the grey color corresponds to the higher
800 inactive compounds (towards the 10,000 nM scale).

801

802 **Figure 8. Feature importance plot from ERα and ERβ models.** Box plots of the top 20
803 features as deduced from the Gini index from RF models built using PubChem fingerprints for
804 both ERα (A) and ERβ (B).

805

806   **Figure 9. Protein structures of the two ER subtypes.** Superimposed structures of ERα (blue)
807   and ERβ (yellow) bound to the E2 ligand (A). Close-up views of the binding cavity of ERα (B)
808   and ERβ (C). Hydrophobic, negatively-charged and positively-charged residues are shown in
809   green, red and blue colored text boxes, respectively.

810

811   **Figure 10. Screenshots of the ERpred web server.** Upon loading of the web server a blank
812   input box (left panel) is shown (A) where users can enter or paste their SMILES notations for use
813   as input for predictions to be made. After predictions are made, results are displayed under the
814   "Status/Output" heading (right panel) (B).

815

Figure 1

Protein structure and sequence alignment of ER subtypes.

ERα and ERβ are displayed in blue and yellow colors, respectively. Secondary structure elements consisting of 12 helices and 2 strands are shown. Sequence alignment of the ligand binding domain of the two ER subtypes were performed in Clustal Omega.

Figure 2

Schematic representation of the methodological workflow of this study.

ERα

11,618
compounds

ERβ

7,810
compounds

ChEMBL
Data Set

Data
Curation

1,593
compounds

Curated
Data Set

1,281
compounds

529
compounds

649
compounds

Active
Set

Inactive
Set

665
compounds

348
compounds

Descriptor
Calculation

Descriptor
Filtering

Balanced
Data Set

Data
splitting
(80/20)

528
compounds

572
compounds

Internal
Set

External
Set

131
compounds

142
compounds

Training

Cross-
validation

External
validation

RF
modeling

Predictive
model

Figure 3

Plot of MW vs ALogP for compounds in the ERα and ERβ datasets.

The plot allows simple visualization of the chemical space of inhibitors against ERα (A) and ERβ (B). Active and inactive compounds are shown in salmon pink and teal colors, respectively.

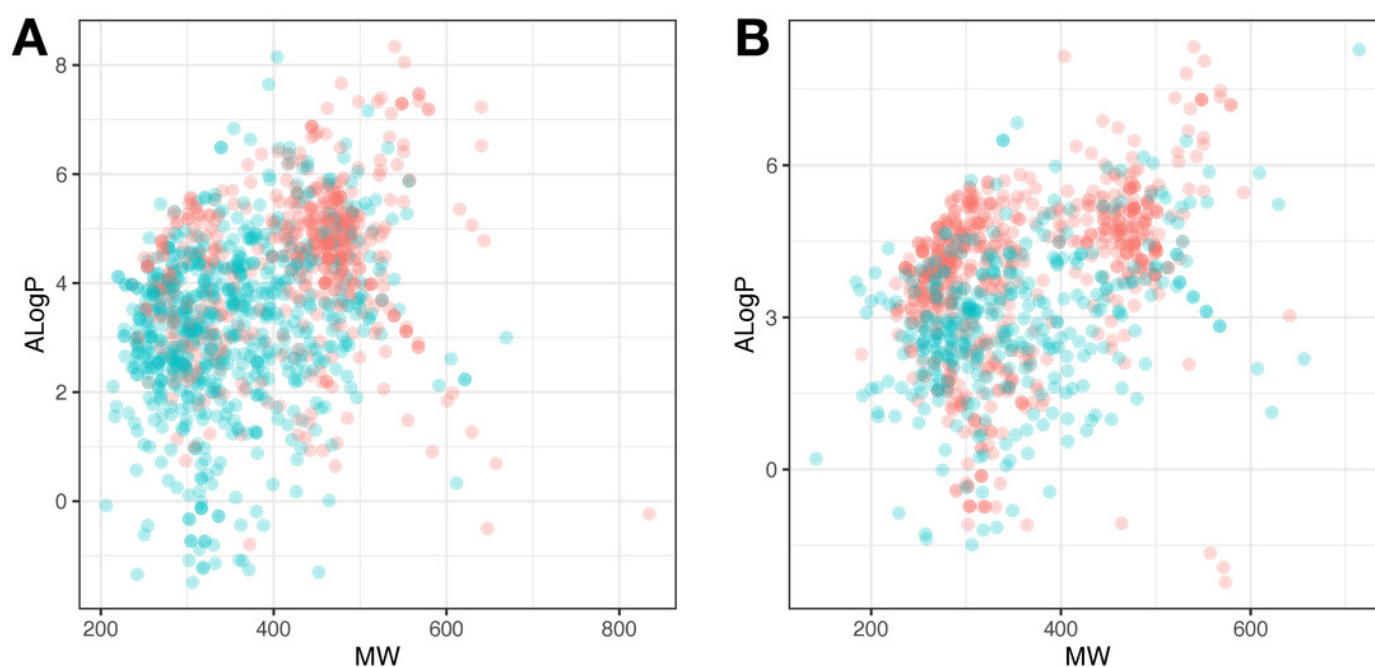



Figure 4

Box plot of Lipinski's rule-of-five descriptors.

The four rule-of-five descriptors are shown for the ERα (A) and ERβ (B) datasets. Active and inactive compounds are shown in salmon pink and teal colors, respectively.

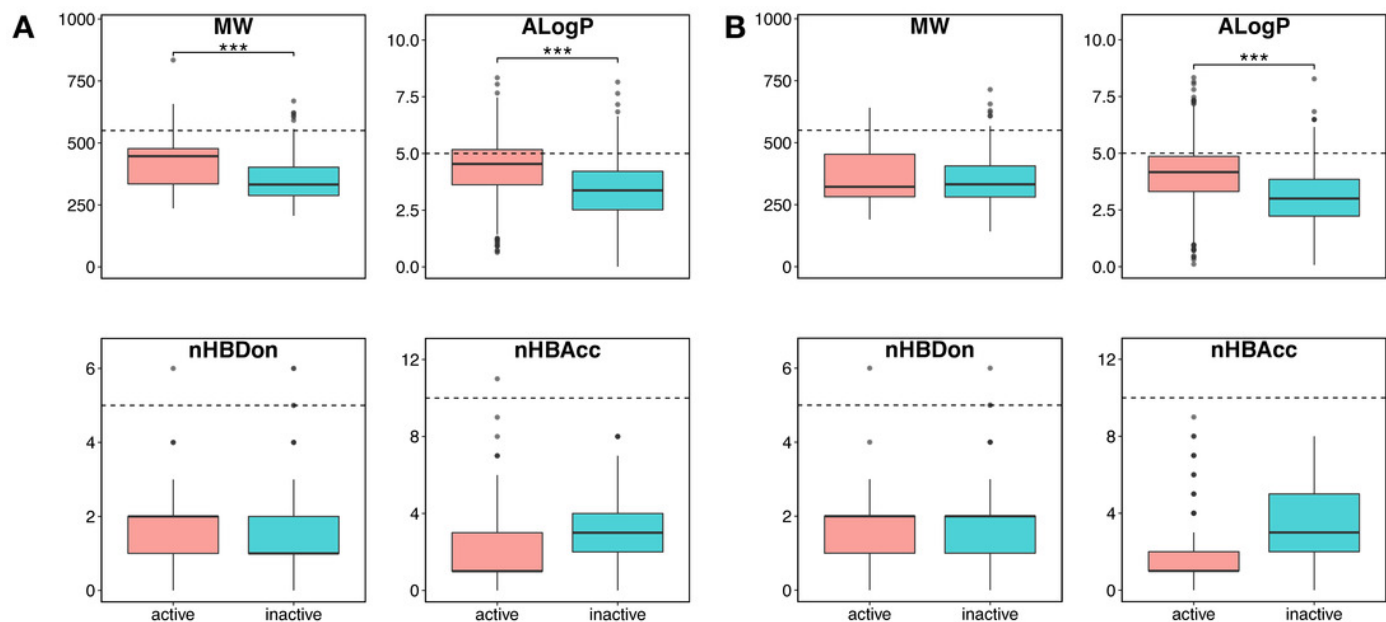



Figure 5

PCA scores plot for compounds in the ERα and ERβ datasets.

The scores plot allows visualization of the distribution of compounds for internal (blue) and external (red) sets that constitutes the assessment of the applicability domain for ERα (A) and ERβ (B) datasets.

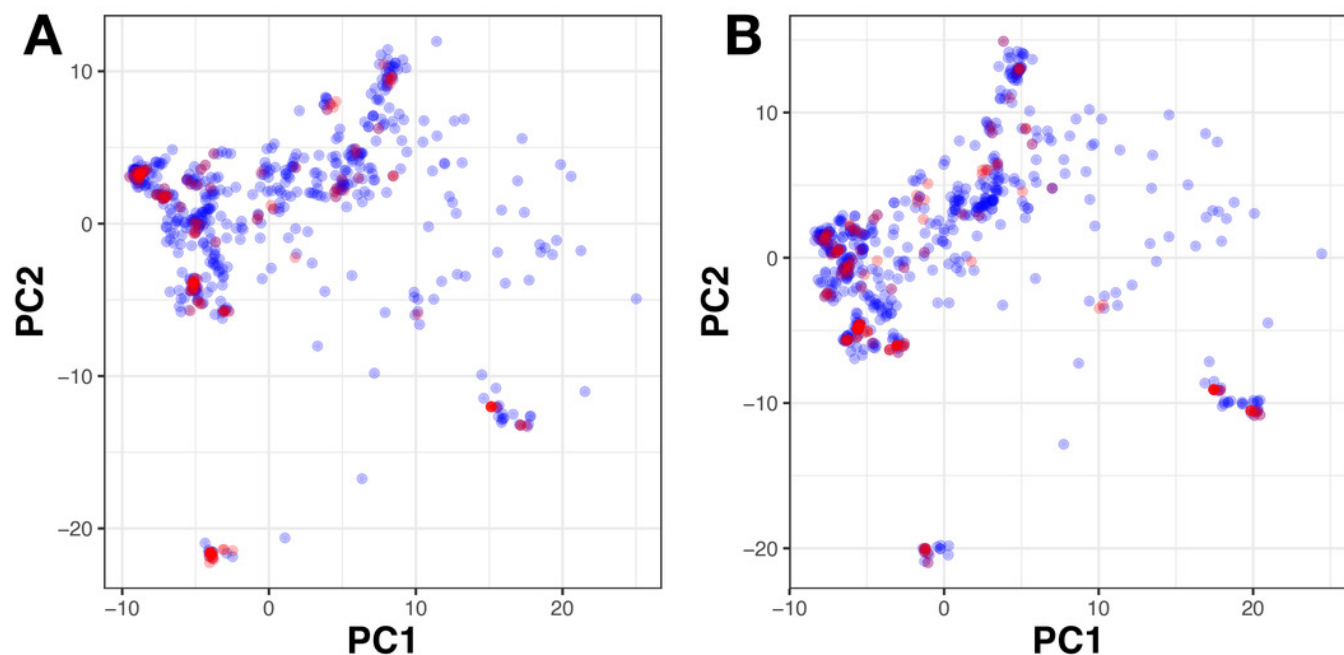



Figure 6

Schematic representation of the methodological workflow of obtaining the scaffolds for ERα and ERβ active and inactive groups.

Top active and inactive compounds were determined from their pIC50 values. Scaffold

Hunter was used to create scaffold trees whereby top scaffolds were determined for each

bioactivity class (i.e., actives and inactives) for both ERα and ERβ.

1. Sort molecules by plC$_{50}$

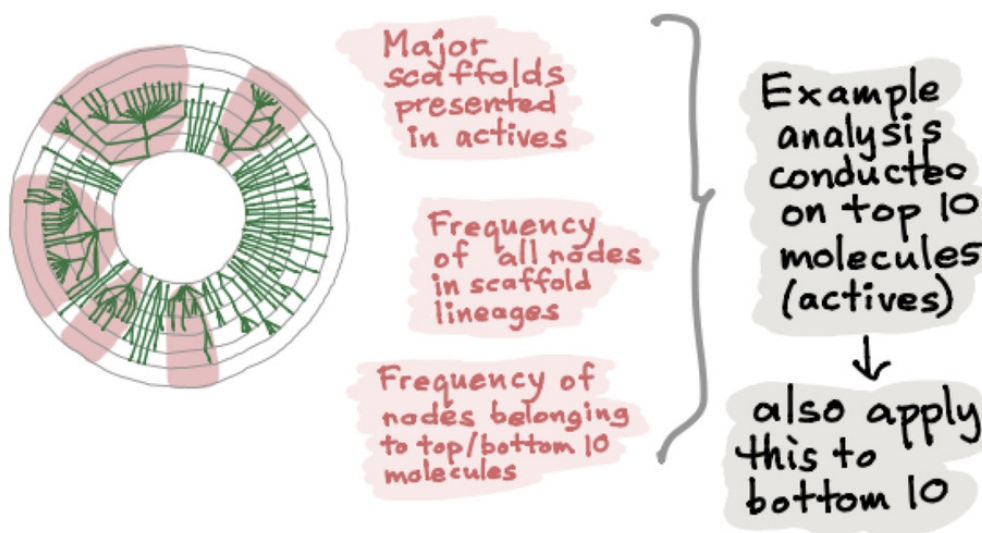
2. Identify common scaffolds from top/bottom 10 molecules

3. Correlate findings with the literature

Figure 7

Chemical structures cloud of ERα and ERβ actives and inactives.

Chemical structure cloud of actives (top panels) and inactives (bottom panels) for ERα and ERβ inhibitors. Particularly, active compounds were defined as compounds having IC50 in the range of 10-1000 nM whereby highly actives (> 10 nM) are represented by salmon pink color while weakly actives (< 10 nM) are represented in grey. Inactive compounds were defined as compounds having IC50 in the range of 10,000-1,000,000 nM whereby the teal color corresponds to the most inactive compounds (towards the 1,000,000 nM scale) while the grey color corresponds to the higher inactive compounds (towards the 10,000 nM scale).

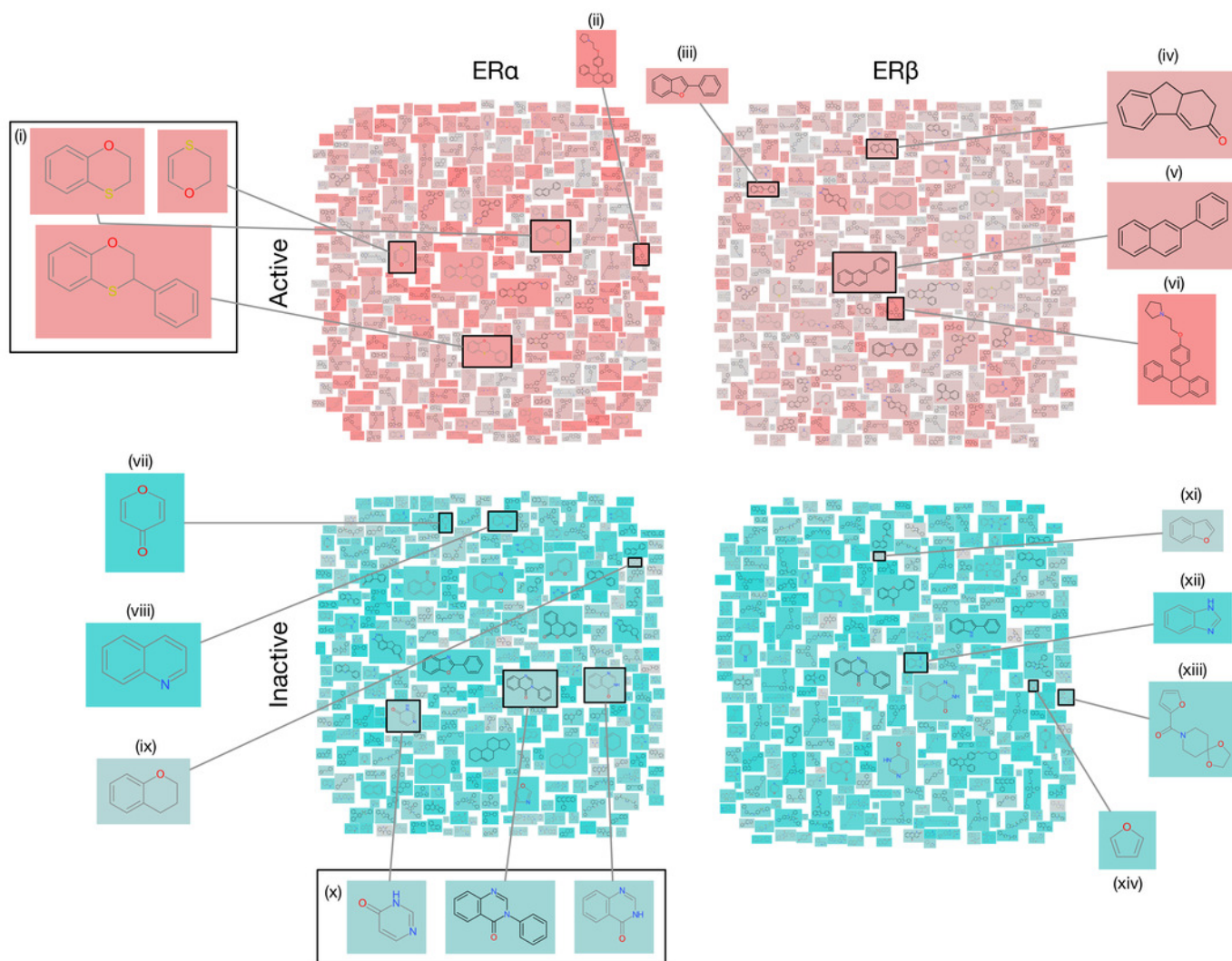
ERα   ERβ

(i)
(ii)
(iii)
(iv)
(v)
(vi)

Active

(vii)
(viii)
(ix)
(x)

Inactive

(xi)
(xii)
(xiii)
(xiv)

(i) 2,3-Dihydro-1,4-benzoxathiine   (ii) diaryltetrahydronaphthalene   (iii) 2-phenylbenzofuran   (iv) 1,2,9,9a-tetrahydrofluoren-3-one   (v) 2-phenylnaphthalene   (vi) diaryltetrahydronaphthalene

(vii) pyran-4-one   (viii) leucoline   (ix) chromane   (x) 4-hydroxypyrimidine and 3-phenyl-4-(3H)quinazolinone   (xi) coumarone   (xii) prazolopyrimidine

(xiii) 1,4-Dioxa-8-azaspiro[4.5]dec-8-yl(2-furyl)methanone   (xiv) furan

Figure 8

Feature importance plot from ERα and ERβ models.

Box plots of the top 20 features as deduced from the Gini index from RF models built using PubChem fingerprints for both ERα (A) and ERβ (B).

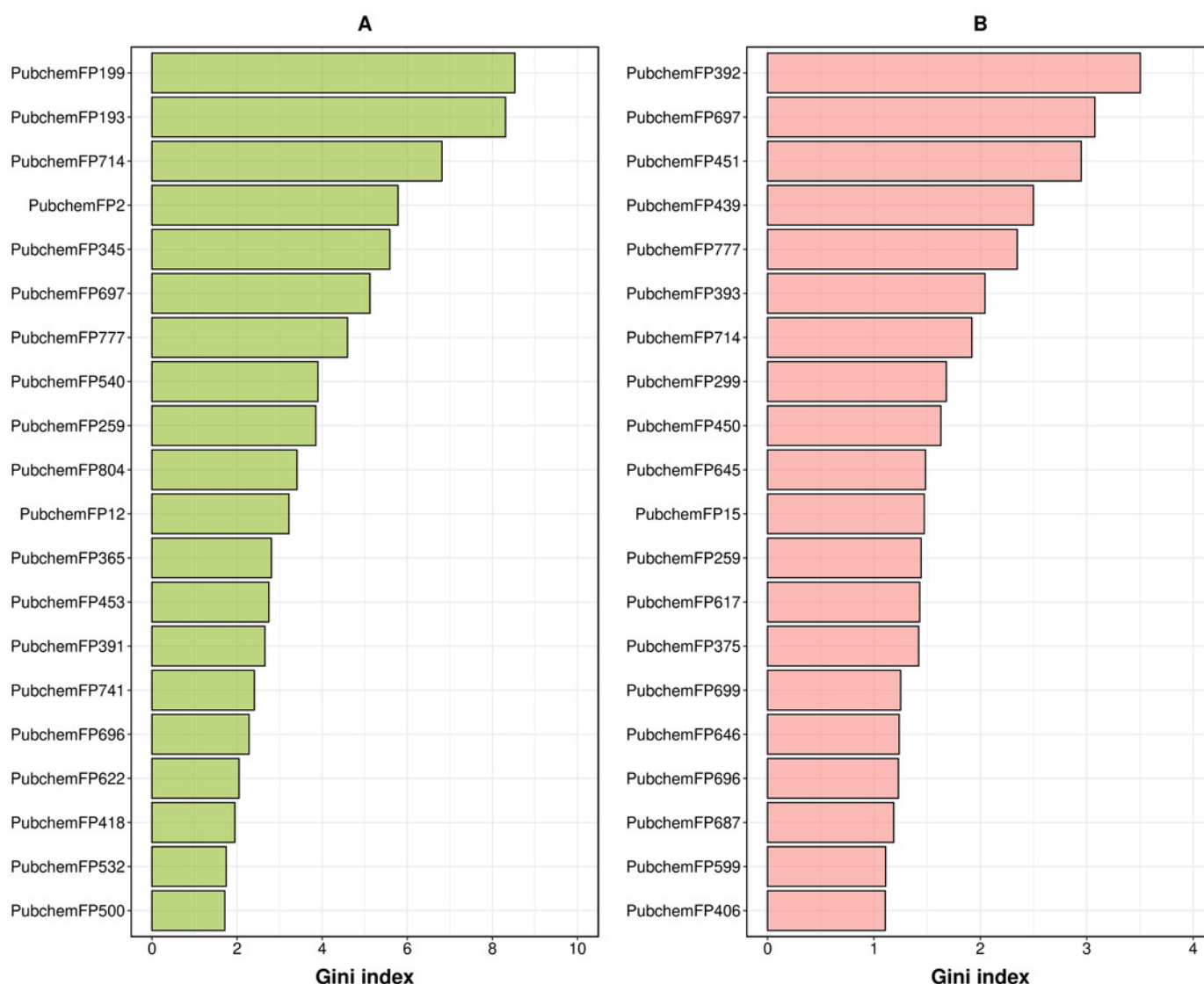



Figure 9

Protein structures of the two ER subtypes.

Superimposed structures of ERα (blue) and ERβ (yellow) bound to the E2 ligand (A). Close-up views of the binding cavity of ERα (B) and ERβ (C). Hydrophobic, negatively-charged and positively-charged residues are shown in green, red and blue colored text boxes, respectively.

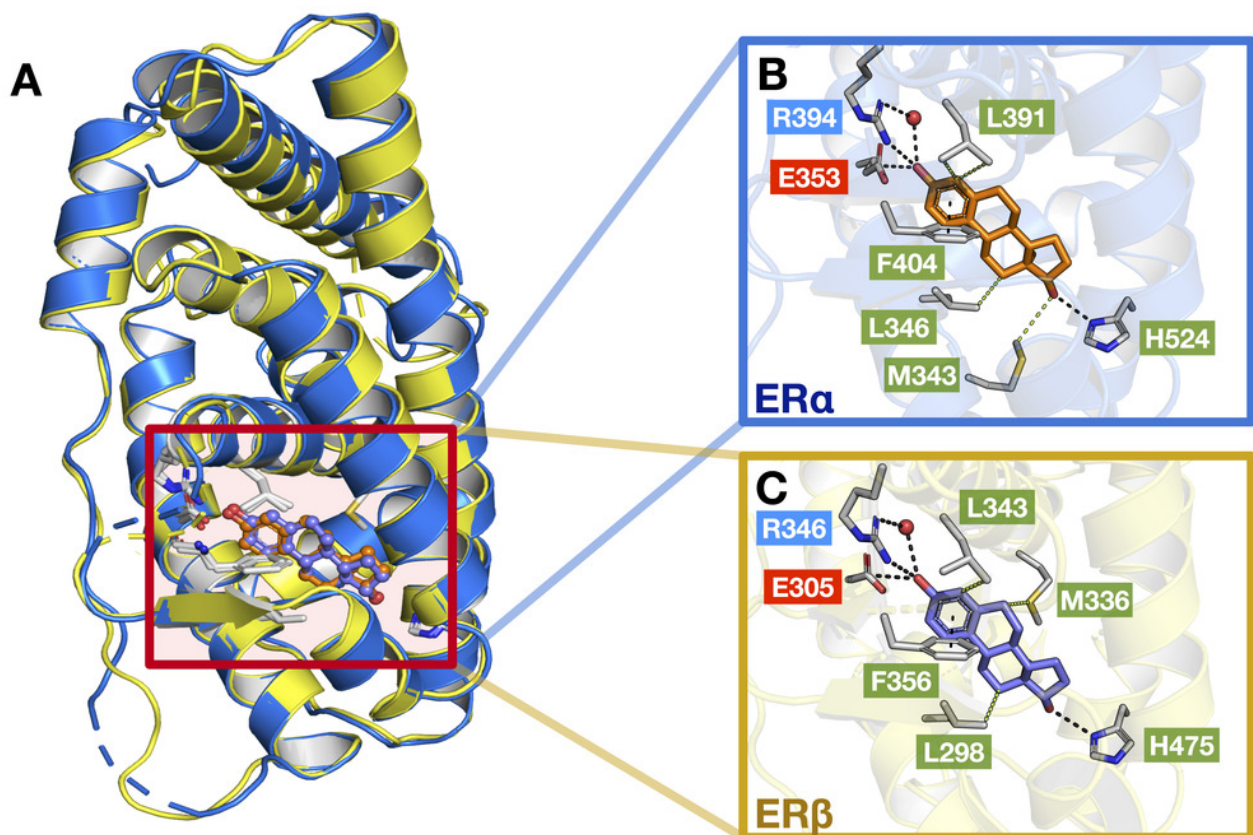



Figure 10

Screenshots of the ERpred web server.

Upon loading of the web server a blank input box (left panel) is shown (A) where users can enter or paste their SMILES notations for use as input for predictions to be made. After predictions are made, results are displayed under the "Status/Output" heading (right panel) (B).

List of 12 sets of fingerprint descriptors calculated from the PaDEL-Descriptor software.

1 **Table 1. List of 12 sets of fingerprint descriptors calculated from the PaDEL-Descriptor**
2 **software.**

Fingerprint	Number	Description
2D Atom Pairs	780	Presence/absence of atom pairs for various topological distances
2D Atom Pairs Count	780	Frequency count of atom pairs for various topological distances
E-state	79	Electrotopological state atom types
CDK	1024	Fingerprint of length 1024 and search depth of 8
CDK Extended	1024	Extends the fingerprint with additional bits describing ring features
CDK Graph Only	1024	A special version considering only the connectivity and not the bond order
Klekota-Roth	4860	Presence/absence of SMARTS patterns for functional groups
Klekota-Roth Count	4860	Frequency count of SMARTS patterns for functional groups
MACCS	166	Binary representation of chemical features defined by Durant et al. (2002) MACCS keys
PubChem	881	Binary representation of substructures as defined by PubChem
Substructure	307	Presence/absence of chemical substructures
Substructure Count	307	Frequency count of chemical substructures

Summary of model performance from classification models for ERα and ERβ.

1 **Table 2. Summary of model performance from classification models for ERα and ERβ.**

Fingerprint	Training set				5-fold cross-validation				Testing set			
	Ac_{Tr}	Sn_{Tr}	Sp_{Tr}	MCC_{Tr}	Ac_{CV}	Sn_{CV}	Sp_{CV}	MCC_{CV}	Ac_{Test}	Sn_{Test}	Sp_{Test}	MCC_{Test}
ERα												
2D Atom Pairs	96.78	95.86	97.90	0.94	83.90	84.14	83.61	0.68	84.73	98.57	68.85	0.72
2D Atom Pairs Count	100.00	100.00	100.00	1.00	87.31	87.50	87.08	0.74	97.71	98.57	96.72	0.95
E-state	89.01	89.47	88.48	0.78	83.33	84.95	81.53	0.67	90.08	94.29	85.25	0.80
CDK	96.59	96.82	96.33	0.93	88.83	87.84	90.09	0.78	93.89	100.00	86.89	0.88
CDK Extended	99.24	99.29	99.18	0.98	89.39	89.00	89.87	0.79	94.66	98.57	90.16	0.89
CDK Graph Only	97.34	97.19	97.53	0.95	86.93	86.64	87.29	0.74	86.26	98.57	72.13	0.74
Klekota-Roth	93.56	95.60	91.37	0.87	86.93	89.34	84.38	0.74	90.84	88.57	93.44	0.82
Klekota-Roth Count	95.83	98.15	93.39	0.92	88.83	91.18	86.33	0.78	86.26	81.43	91.80	0.73
MACCS	96.40	96.15	96.69	0.93	84.66	85.31	83.88	0.69	96.95	95.71	98.36	0.94
PubChem	96.40	96.15	96.69	0.93	87.69	86.09	89.82	0.75	94.66	100.00	88.52	0.90
Substructure	92.99	93.93	91.94	0.86	83.52	86.57	80.38	0.67	93.89	97.14	90.16	0.88
Substructure Count	91.28	93.09	89.33	0.83	82.95	85.87	79.92	0.66	93.89	97.14	90.16	0.88
ERβ												
2D Atom Pairs	96.68	96.32	98.18	0.90	86.01	87.53	77.11	0.55	88.73	99.10	51.61	0.65
2D Atom Pairs Count	99.65	99.55	100.00	0.99	88.46	90.11	80.41	0.64	92.96	100.00	67.74	0.79
E-state	94.06	94.03	94.17	0.82	87.24	89.45	76.53	0.60	90.14	98.20	61.29	0.69
CDK	99.13	99.11	99.18	0.97	90.03	91.14	84.69	0.69	95.77	98.20	87.10	0.87
CDK Extended	98.95	98.89	99.17	0.97	90.73	92.09	84.62	0.72	94.37	99.10	77.42	0.83
CDK Graph Only	96.85	96.53	98.20	0.91	87.41	89.31	77.89	0.61	91.55	98.20	67.74	0.74
Klekota-Roth	98.25	98.23	98.32	0.95	89.16	90.87	81.19	0.66	94.37	99.10	77.42	0.83
Klekota-Roth Count	98.95	98.89	99.17	0.97	90.38	91.53	85.00	0.70	94.37	99.10	77.42	0.83
MACCS	99.48	99.78	98.41	0.98	88.81	91.18	78.50	0.66	95.07	99.10	80.65	0.85
PubChem	98.25	98.02	99.15	0.95	90.38	91.53	85.00	0.70	92.25	99.10	67.74	0.76
Substructure	95.10	94.67	97.09	0.85	87.94	89.71	79.17	0.62	94.37	99.10	77.42	0.83
Substructure Count	99.30	99.33	99.19	0.98	89.34	91.24	80.77	0.67	95.77	99.10	83.87	0.87

Summary of the top 20 features from the ERα model along with their corresponding SMARTS patterns and description.

The top features were obtained from the feature importance plot of the RF model.

1   **Table 3. Summary of the top 20 features from the ERα model along with their**
2   **corresponding SMARTS patterns and description.** The top features were obtained from the
3   feature importance plot of the RF model.

Features	SMARTS pattern	Substructure description
PubChemFP199	>= 4 any ring size 6	Greater than or equal to 4 six-membered cyclic ring
PubChemFP193	>= 3 saturated or aromatic carbon-only ring size 6	Greater than or equal to 3 saturated or aromatic carbon-only six-membered cyclic ring
PubChemFP714	Cc1ccc(O)cc1	4-methylphenol
PubChemFP2	>= 16 H	Greater than or equal to sixteen hydrogen atoms
PubChemFP345	C(~C)(~H)(~N)	Ethylamine
PubChemFP697	C-C-C-C-C-C(C)-C	2-methylheptane
PubChemFP777	CC1CCC(O)CC1	4-methylphenol
PubChemFP540	C-N-C-[#1]	1-(2-chloroethyl)-3-[2-[2-[[2-chloroethyl(nitroso)carbamoyl]amino]ethyl disulfanyl]ethyl]-1-nitrosourea
PubChemFP259	>= 3 aromatic rings	Greater than or equal to 3 aromatic rings
PubChemFP804	OC1CC(S)CCC1	3-sulfonyl phenol
PubChemFP12	>= 16 C	Greater than or equal to sixteen carbon atoms
PubChemFP365	C(~H)(~N)	Methanamine
PubChemFP453	N(-C)(=C)	N-methylmethanimine
PubChemFP391	N(~C)(~C)(~C)	N,N-dimethylmethanamine
PubChemFP741	Oc1cc(S)ccc1	3-sulfonyl phenol
PubChemFP696	C-C-C-C-C-C-C-C	Octane
PubChemFP622	O=C-O-C:C	Ethyl formate
PubChemFP418	C=N	Methanimine
PubChemFP532	S-C:C-[#1]	Ethanethiol
PubChemFP500	C-S-C:C	Methylsulfanylethane

Summary of top 20 features from the ERβ model along with their corresponding SMARTS patterns and description.

The top features were obtained from the feature importance plot of the RF model.

1 **Table 4. Summary of top 20 features from the ERβ model along with their corresponding**
2 **SMARTS patterns and description.** The top features were obtained from the feature
3 importance plot of the RF model.

Features	SMARTS pattern	Substructure description
PubChemFP392	N(~C)(~C)(~H)	*N*-methylmethanamine
PubChemFP697	C-C-C-C-C-C(C)-C	2-methylheptane
PubChemFP451	C(-N)(=O)	Formamide
PubChemFP439	C(-C)(-N)(=O)	Acetamide
PubChemFP777	CC1CCC(O)CC1	4-methylphenol
PubChemFP393	N(~C)(~H)	Methanamine
PubChemFP714	Cc1ccc(O)cc1	4-methylphenol
PubChemFP299	N-H	Lambda1-azane
PubChemFP450	C(-N)(=N)	Methanimidamide
PubChemFP645	O=C-N-C-C	N-ethylformamide
PubChemFP15	>= 2 N	Greater than or equal to two nitrogen atoms
PubChemFP259	>= 3 aromatic rings	Greater than or equal to three aromatic rings
PubChemFP617	C-C-C-O-[#1]	Propan-1-ol
PubChemFP375	C(~N)(~N)	Methanediamine
PubChemFP699	O-C-C-C-C-C(C)-C	5-methylhexan-1-ol
PubChemFP646	O=C-N-C-[#1]	*N*-methylformamide
PubChemFP696	C-C-C-C-C-C-C-C	Octane
PubChemFP687	O=C-C-C-C=O	Butanedial
PubChemFP599	[#1]-C-C=C-[#1]	Prop-1-ene
PubChemFP406	O(~C)(~H)	Methanol