INSTITUTO POLITÉCNICO DE BRAGANÇA
Escola Superior Agrária

QSAR modeling studies of a library of Human Tyrosinase inhibitors

Cristiano Gabi dos Santos Mateus

Dissertation submitted to Escola Superior Agrária de Bragança to obtain the Degree of Master in Biotechonological Engineering

Supervised by

Professor Doutor Rui Miguel Vaz de Abreu

Doutora Lillian Bouçada de Barros

This dissertation does not include the comments and suggestions mentioned by the Jury

Bragança
2022

Agradecimentos

Primeiramente gostaria de agradecer aos meus orientadores, Professor Doutor Rui Miguel Vaz de Abreu e Doutora Lillian Bouçada de Barros, por me terem aceitado como orientando e por me darem um voto de confiança para trabalhar numa área que me deu e dá muito gosto em trabalhar e na qual pretendo alargar os meus conhecimentos. Em especial ao Professor Doutor Rui Abreu gostaria de dizer que foi uma ótima experiência e um enorme privilégio trabalhar consigo. O meu muito obrigado pelo incentivo, dedicação, ideias, sugestões e por me transmitir todos os seus conhecimentos nesta área, fazendo com que ambicione mais e queira realizar mais trabalhos que envolvam bioinformática.

Deixo também o meu muito obrigado a todos os que retiraram do seu tempo para me aconselharem e transmitirem os seus conhecimentos que permitiram resolver muitos problemas que foram surgindo durante a elaboração desta dissertação, nomeadamente, o meu muito obrigado a: Professor Doutor Luís Dias, Professor Doutor José Rufino, Doutor Bruno Melgar, Doutor José Ignacio, Doutor Taofiq Oludemi e doutorando Carlos Shiraishi.

Quero dar o meu enorme agradecimento à minha família, em especial aos meus pais e avós pois, sem eles, este capítulo da minha vida não se estaria a realizar. Obrigado por sempre me apoiarem e me darem toda a ajuda para que este dia se concretizasse. À minha namorada, deixo o meu agradecimento por me acompanhar em mais uma etapa importante da minha vida. Obrigado por todo o carinho, apoio, compreensão e paciência que tiveste e continuas a ter para comigo.

Por último, mas não menos importante, deixo o meu agradecimento aos meus amigos e amigas que sempre me aconselharam, apoiaram e incentivaram. Em especial, o meu obrigado a: Maria do Céu Fidalgo, Ana Sagres, Sofia Silva, Júlia Machado, Filipe Lema, Diogo Félix, Bruna Carbas, Ana Rodrigues e Nuno Ferreiro.

A todos, o meu sincero obrigado!

Index

Index of Figures

Index of tables and graphs

List of abbreviations

TYR – Tyrosinase

TYRP1 – Tyrosinase-related protein-1

DCT – DOPAchrome tautomerase

Pmel17 – Premelanosome protein 17

MART1 – Melanoma antigen recognized by T cells-1

AP-3 – Adaptor protein complex 3

BLOC-1 – Biogenesis of lysosome-related organelles complex 1

OCA2 – P protein

TYRP2 – Tyrosinase-related protein-2

MITF – Microphthalmia-associated transcription factor

DOPA – Dihydroxyphenylalanine

DHI – 5,6-dihydroxyindole

DHICA – 5,6-dihydroxy-1H-indole-2-carboxylic acid

***hs*TYR** – Human tyrosinase

UV – Ultraviolet

SNpc – Substantia nigra pars compacta

HGD – Homogentisate 1,2-dioxygenase

HGA – Homogentisic acid

***ab*TYR** – *Agaricus bisporus* tyrosinase

QSAR – Quantitative structure-activity relationship

GA – Genetic algorithm

MLR – Multiple linear regression

R^2 – Determination coefficient

Abstract

Melanogenesis is the chemical process responsible for synthesizing melanin, which occurs in melanocytes, in subcellular lysosome-like organelles called melanosomes. Melanin plays a vital role in protecting the skin from damage caused by ultraviolet rays. However, excess melanin production or abnormal distribution can cause various pigmentation disorders, such as over-tanning, age spots, and melasma. Skin disorders like these, have prompted the development of skin-whitening compounds to reduce melanin content. Furthermore, inhibition of melanin synthesis is considered a valid therapeutic strategy for treating advanced melanotic melanomas

Human tyrosinase (*hs*TYR) is the most important enzyme involved in the melanogenesis process, as it catalyzes, at least, its first two steps. Tyrosinase from the white button mushroom *Agaricus bisporus* (*ab*TYR) has been widely available at low cost from commercial sources for several decades, whereas *hs*TYR is still expensive and difficult to produce. The importance of discovering more and better *hs*TYR inhibitors has been widely discussed, as when tested against *hs*TYR, several *ab*TYR inhibitors provide disappointing results, including some of the most extensively used depigmenting compounds now used in dermocosmetics.

An *in silico* methodology that can be used to predict compound bioactivities is QSAR (quantitative structure-activity relationship) modelling. A QSAR model tries to find correlations between a biological activity of interest and molecular descriptors calculated from the compound structure. In this work, a QSAR model was developed to predict *hs*TYR inhibition activity using the PYTHON computer language and its PyQSAR package. To develop a QSAR model, a library of 196 known *hs*TYR inhibitors was gathered, and compounds were divided into 6 groups according to their scaffold structure. A total of 33 QSAR models were prepared using different combinations of the defined groups and different pools of molecular descriptors.

QSAR model 32 was selected for further use as it presented good statistical robustness and had the highest number of compounds, 41 in total. Of the 28,933 molecular descriptors calculated by the OCHEM platform for the 41 compounds used, PyQSAR selected 4 to be used in the model: C-026; DISSM2C; MaxdssC; WHALES90_Rem. The statistical data obtained after the validation of the QSAR model by cross-validation was excellent, namely the determination coefficient ($R^2CV=0.9147$), the value of the square

root of the mean error (RMSE CV=0.1878) and the mean value of the score of the multiple linear regression method (Q^2CV=0.8922). This QSAR model originates a mathematical equation that allows the prediction of *hs*TYR inhibition activity by new compounds with similar structures.

A library of natural compounds, with a structure similar to those used to develop QSAR model 32, was created using the COCONUT database of natural compounds. A total of 1,628 natural compounds were gathered, their molecular descriptors were calculated, and the QSAR model 32 equation was applied. The results are displayed on a website and can be viewed by accessing the URL http://esa.ipb.pt/qsar/. The ZINC15 database was used to determine which of the compounds in the developed natural compound library would be available for purchase after predicting the *hs*TYR inhibitory activity of each compound in the library. A total of 18 different compounds were bought from different companies. To evaluate these compounds experimental ability to inhibit *hs*TYR and thus validate QSAR model 32, the compounds will be tested against this enzyme. If those compounds activity is confirmed, they may be used in cosmeceutical applications.

Keywords: QSAR, PYTHON, PyQSAR, molecular descriptor, melanin, *hs*TYR, *ab*TYR, OCHEM, COCONUT, ZINC15.

Resumo

A melanogénese é o processo químico responsável pela síntese da melanina, que ocorre nos melanócitos, em organelos subcelulares semelhantes aos lisossomas chamados melanossomas. A melanina desempenha um papel vital na proteção da pele dos danos causados pelos raios ultravioleta. No entanto, a produção excessiva de melanina ou distribuição anormal pode causar vários distúrbios de pigmentação, como bronzeamento excessivo, manchas senis e melasma. Distúrbios de pele como estes levaram ao desenvolvimento de compostos de clareamento da pele para reduzir o conteúdo de melanina. Além disso, a inibição da síntese de melanina é considerada uma estratégia terapêutica válida para o tratamento de melanomas melanóticos avançados

A tirosinase humana (*hs*TYR) é a enzima mais importante envolvida no processo de melanogénese, pois catalisa, pelo menos, as suas duas primeiras etapas. A tirosinase do cogumelo branco *Agaricus bisporus* (*ab*TYR) está amplamente disponível a baixo custo em fontes comerciais há várias décadas, enquanto a *hs*TYR ainda é cara e difícil de produzir. A importância de descobrir mais e melhores inibidores de *hs*TYR tem sido amplamente discutida, pois quando testados contra *hs*TYR, vários inibidores de *ab*TYR fornecem resultados dececionantes, incluindo alguns dos compostos despigmentantes mais usados atualmente em dermocosméticos.

Uma metodologia *in silico* que pode ser usada para prever bioatividades compostas é a modelação QSAR (quantitative structure-activity relationship). Um modelo QSAR tenta encontrar correlações entre uma atividade biológica de interesse e descritores moleculares calculados a partir da estrutura do composto. Neste trabalho, um modelo QSAR foi desenvolvido para prever a atividade de inibição de *hs*TYR usando a linguagem de computador PYTHON e seu pacote PyQSAR. Para desenvolver um modelo QSAR, uma biblioteca de 196 inibidores *hs*TYR conhecidos foi reunida e os compostos foram divididos em 6 grupos de acordo com sua estrutura de base. Um total de 33 modelos QSAR foram preparados usando diferentes combinações dos grupos definidos e diferentes pools de descritores moleculares.

O modelo QSAR 32 foi selecionado para uso posterior por apresentar boa robustez estatística e possuir o maior número de compostos, 41 no total. Dos 28 933 descritores moleculares calculados pela plataforma OCHEM para os 41 compostos utilizados, o PyQSAR selecionou 4 para serem utilizados no modelo: C-026; DISSM2C; MaxdssC;

WHALES90_Rem. Os dados estatísticos obtidos após a validação do modelo QSAR por validação cruzada foram excelentes, nomeadamente o coeficiente de correlação (R^2CV=0,9147), o valor da raiz quadrada do erro médio (RMSE CV=0,1878) e o valor médio da pontuação do método de regressão linear múltipla (Q^2CV=0,8922). Este modelo QSAR origina uma equação matemática que permite prever a atividade de inibição de *hs*TYR por novos compostos com estruturas semelhantes.

Uma biblioteca de compostos naturais, com uma estrutura similar àquelas usadas para desenvolver o modelo QSAR 32, foi criada usando o banco de dados de compostos naturais COCONUT. Um total de 1 628 compostos naturais foram recolhidos, os seus descritores moleculares calculados e a equação do modelo QSAR 32 foi aplicada. Os resultados são apresentados num website criado por nós e podem ser visualizados acedendo ao URL http://esa.ipb.pt/qsar/. O banco de dados ZINC15 foi usado para determinar quais compostos na biblioteca de compostos naturais desenvolvidos estariam disponíveis para compra após prever a atividade inibitória de *hs*TYR de cada composto na biblioteca. Um total de 18 compostos diferentes foram comprados de diferentes empresas. Para avaliar a capacidade experimental destes compostos em inibir a *hs*TYR e assim validar o modelo QSAR 32, os compostos serão testados contra esta enzima. Caso a atividade desses compostos seja confirmada, eles poderão ser utilizados em aplicações cosmecêuticas.

Palavras-chave: QSAR, PYTHON, PyQSAR, descritor molecular, melanina, *hs*TYR, *ab*TYR, OCHEM, COCONUT, ZINC15.

1. Introduction

1.1 Melanogenesis

Melanogenesis is the chemical process responsible for the synthesis of melanin, which occurs in melanocytes, in subcellular lysosome-like organelles called melanosomes.[1] These intracellular corpuscles are specialized organelles of pigmented cells that are responsible for the synthesis, storage, and transport of melanin pigments, which are responsible for most visible pigmentation in mammals and other vertebrates.[2] Melanosomes require a number of specific enzymatic and structural proteins to mature and become competent in order to produce melanin. Tyrosinase (TYR), tyrosinase-related protein-1 (TYRP1), and DOPAchrome tautomerase (DCT) are among the critical enzymes that affect melanin quantity and quality, whilst Pmel17 (premelanosome protein 17) and MART1 (melanoma antigen recognized by T cells-1) are critical structural proteins required for the structural maturation of melanosomes.[3] AP-3 (adaptor protein complex 3), BLOC-1 (Biogenesis of lysosome-related organelles complex 1) and OCA2 (P protein)[4] have important roles in sorting and trafficking melanosomes.

Melanocytes are dendritic cells of the neuroectoderm, and their primary function is the production of melanin pigment (**Figure 1**).[5]

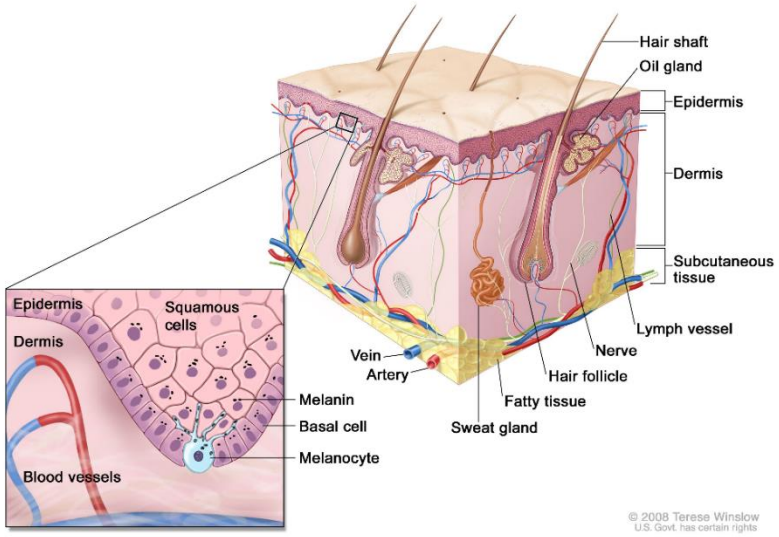


Figure 1 – Illustration of the epidermis, dermis, and subcutaneous tissue in the skin. The layer of basal cells at the deepest section of the epidermis contains melanocytes.[6]

The precursor cells of melanocytes, called melanoblasts, are unpigmented cells that arise from embryonic neural crest cells. After neural tube closure, melanoblasts migrate

to various regions of the body and develop into melanocytes as well as peripheral nervous system cells, bone, and cartilage of the head and choroid of the eye.[5] Melanoblasts that develop into melanocytes are found predominantly in the skin epidermis basal layer and hair follicles and can be identified by the expression of specific melanocyte markers such as: TYR, TYRP1, TYRP2 (tyrosinase-related protein-2), DCT, Pmel17, MART-1 and MITF (microphthalmia-associated transcription factor).[7]

Usually, there are two types of melanin being produced: pheomelanin (showing tones between yellow to red) and eumelanin (showing shades between brown to black). The difference between these two types of melanin is in their chemical structure and synthesis pathway. The type of melanin produced is determined by the availability of substrates and the function of the enzymes involved in melanogenesis.

The simplified metabolic pathway for the synthesis of these two compounds is shown in **Figure 2**. TYR is the main enzyme in this pathway, first promoting Tyrosine hydroxylation into DOPA (dihydroxyphenylalanine) and then oxidation of DOPA to DOPAquinone. In the presence of cysteine, DOPAquinone is then oxidized and polymerized to form pheomelanin. When cysteine is found in lower concentrations, the eumelanin production pathway is activated.[8] In this case, DOPAquinone is transformed into DOPAchrome, which can spontaneously lose a carboxylic group to form 5,6-dihydroxyindole (DHI). DOPAchrome can, with the help of DOPAchrome tautomerase, or TYRP2, be tautomerized and originate 5,6-dihydroxy-1H-indole-2-carboxylic acid (DHICA). Both DHI and DHICA can be further oxidized and polymerized to form a high molecular density complex known as DHI-melanin, which is then transformed into eumelanin.

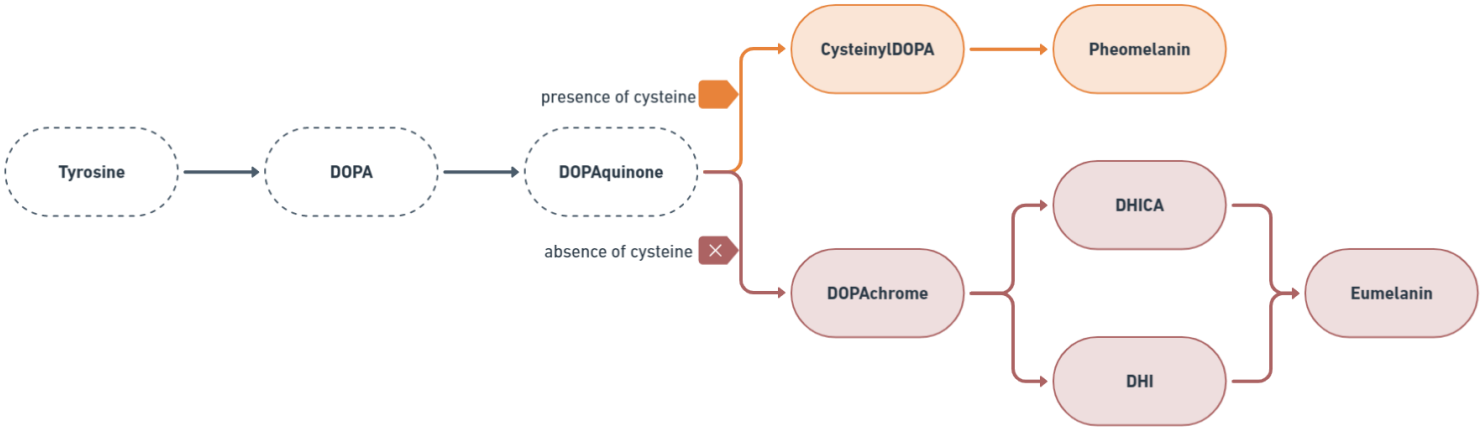


Figure 2 – Simplified scheme of the metabolic pathway of two types of melanin: pheomelanin and eumelanin.

1.2 Human tyrosinase

Tyrosinases, also known as monophenol or *o*-diphenol:oxygen oxidoreductase (EC:1.14.18.1), are type 3 copper proteins with two copper ions in the active site. As it's shown in **Figure 3**, these enzymes catalyze the conversion of monophenols like tyrosine into *o*-diphenols, followed by the oxidation of the *o*-diphenols to the corresponding *o*-quinone derivatives. The related catechol oxidases only catalyze the second reaction, using *o*-diphenols as substrates.[9]

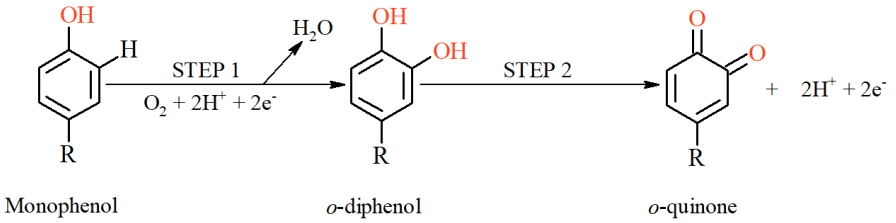


Figure 3 – Generic actions of monophenol monooxygenase enzyme.[9]

Human tyrosinase (*hs*TYR) is the most important enzyme involved in the melanogenesis process, as it catalyzes, at least, the first two steps of this process. This glycoprotein has a molecular weight of 67 kDa and is composed of 529 amino acids, including an 18-residue N-terminal signal sequence and six or seven N-glycosylated positions. Its active site is made up of two close, magnetically coupled copper centers that are connected by an aquo(hydroxo) ligand in the met state (*hs*TYR's reactive state) and coordinated by six histidine residues (H180, H202, H211 for CuA, H363, H367, H390 for CuB), which are highly conserved among tyrosinases, catechol oxidases, and hemocyanins (**Figure 4**).[10]

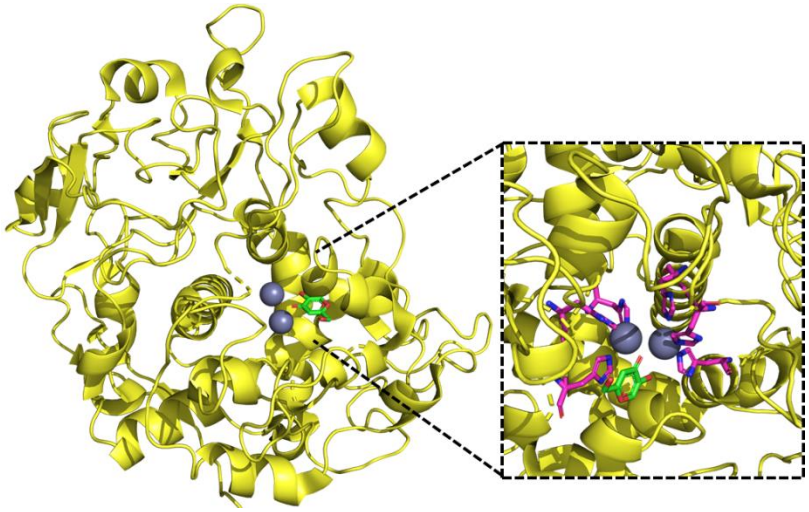


Figure 4 – 3D structure of *hs*TYR. In this image it is possible to observe the active site of the enzyme composed of copper centers (gray spheres) where a molecule of kojic acid is attached (green structure). It is possible, through a close-up of the image, to observe the 6 histidine residues (pink structures) surrounding the copper centers.

Although the enzyme can oxidize a wide range of monophenolic and diphenolic substrates, the physiological function of *hs*TYR is to o-hydroxylate L-tyrosine to L-DOPA (monophenolase activity) and then oxidize L-DOPA in DOPAquinone (diphenolase Activity) using molecular oxygen (**Figure 5**).[11]

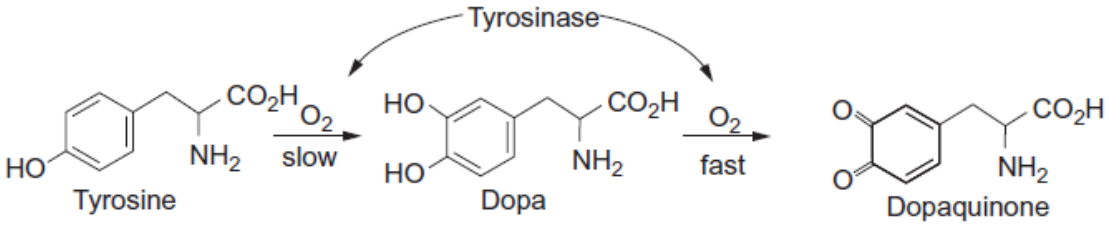


Figure 5 – Tyrosinase actions in the first two stages of melanogenesis.[9]

This double oxidation process starts the synthesis of melanin pigments, which is mostly nonenzymatic afterwards. Only two other enzymes are known to be involved in melanogenesis, *hs*TYRP1 and *hs*TYRP2, which share a high level of homology with *hs*TYR. However, the role of *hs*TYRP1 is unknown and may be linked to *hs*TYR protection.[12]

In **Figure 6** it is possible to observe the complete metabolic pathway for the synthesis of the different melanin components and the importance of *hs*TYR. While the metabolic pathways of eumelanin and pheomelanin are well known, the existence of neuromelanin is still debated. Still, the current scientific consensus is that it exists in the brain as a dark pigment comprised of a polymer of 5,6-dihydroxyindoles.[13] Thus, in addition to its role in peripheral (cutaneous and ocular) melanogenesis, *hs*TYR may play a significant role in neuromelanin synthesis in the brain. However, because only trace levels of *hs*TYR mRNA and *hs*TYR itself were found in the human brain, the enzyme's role in neuromelanin synthesis has long been discussed.[10]

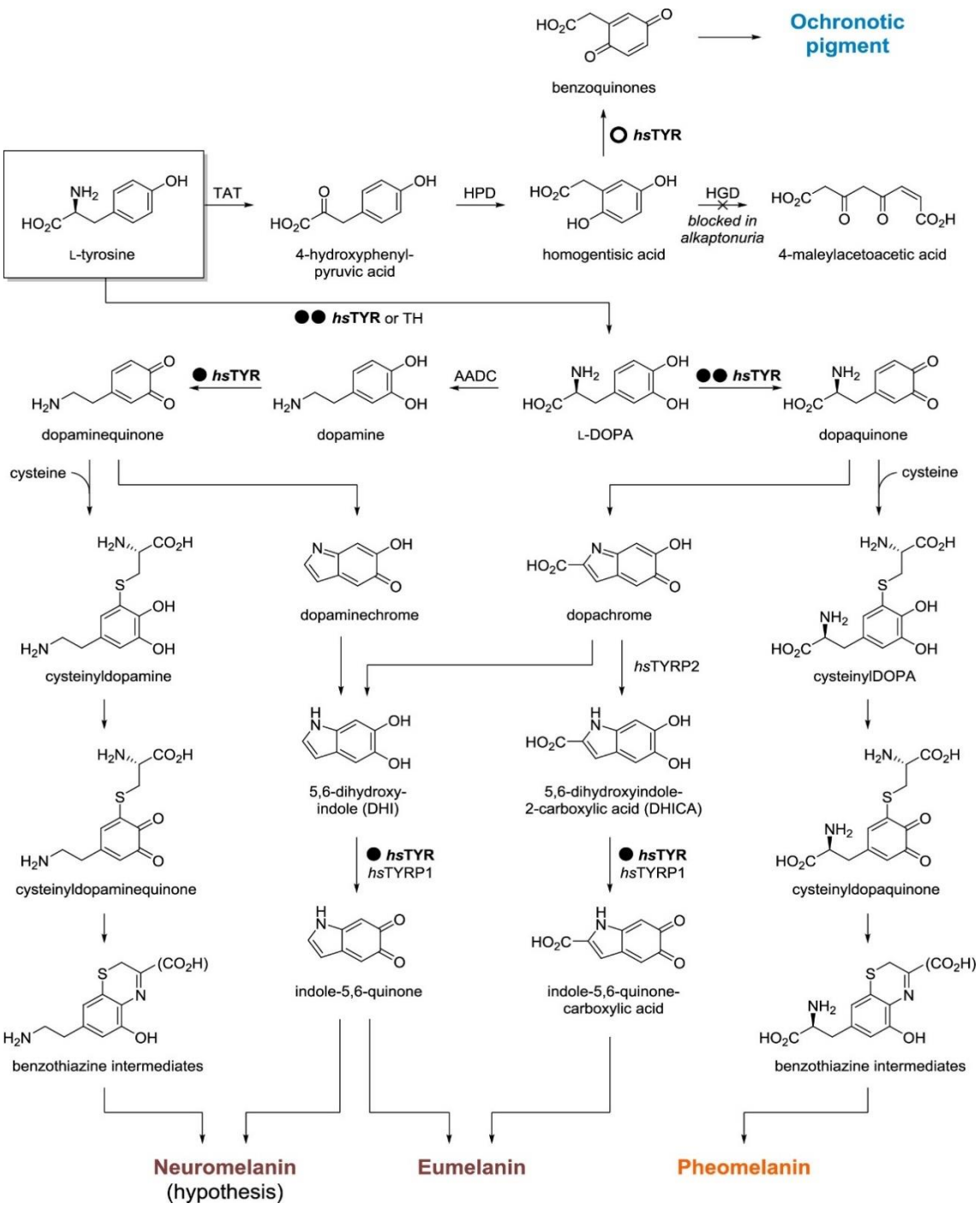


Figure 6 – L-Tyrosine-related pigmentation biochemical pathways involving *hs*TYR as a demonstrated (●●), suspected (●) or purely hypothetic (○) component.[10] TAT – Tyrosine aminotransferase (EC no. 2.6.1.5); HPD – 4-Hydroxyphenylpyruvate dioxygenase (EC no. 1.13.11.27); HGD – Homogentisate 1,2-dioxygenase (EC no. 1.13.11.5); TH – Tyrosine hydroxylase (EC no. 1.14.16.2).

1.3 Benefits of inhibiting melanogenesis

Melanin plays a significant role in protecting the skin from damage caused by ultraviolet (UV) rays, and this property is due to melanin's ability to absorb free radicals generated within the cytoplasm of cells by the action of UV rays.[14] However, excess

melanin production or abnormal distribution can cause pigmentation disorders, such as over-tanning, age spots, and melasma.[15] These skin disorders have prompted the development of skin-whitening compounds to reduce melanin content. Furthermore, inhibition of melanin synthesis is considered a valid therapeutic strategy for treating advanced melanotic melanomas.[16]

As previously stated, the enzyme *hs*TYR plays a critical role in the melanogenic process since it catalyzes the rate-limiting first two steps of the reaction sequence (**Figure 6**). As a result, most efforts to suppress or reduce melanogenesis *in vivo* have been devoted to discovering *hs*TYR inhibitors. The strategy has proven effective, with clear correlations observed between the level of pigmentation produced and the inhibition of *hs*TYR. As an outcome, *hs*TYR appears to be a convenient and appealing target for reducing human melanogenesis in various contexts.

1.3.1 Skin-whitening agents: dermatological and cosmetic applications

The most obvious and requested application for melanogenesis inhibition is skin whitening. This practice, prevalent in some ethnic groups, particularly in Asia, Africa, and the Middle East, is the result of a complex interplay of cultural, social, political, and psychological factors, and it has been documented since the dawn of human civilizations.[17] In these cultures, lighter skin tones are often associated with beauty and health, whereas darker skin tones are associated with lower social status, and thus interest in skin whitening has grown exponentially since the 1980s.[18] A study by Sagoe et al. (2019)[19] revealed that about 28% of the global population practices skin-whitening at least once in a lifetime, suggesting that this market has enormous potential for growth, progression, and profitability. Despite its aesthetic value, depigmentation is also a medical necessity for patients suffering from common to extremely rare dermatological disorders such as melasma, solar lentigo, erythromelanosis follicularis faciei et colli, erythema dyschromicum perstans, congenital melanocytic naevi and postinflammatory hyperpigmentation. These dermatological disorders are always associated with hyperpigmentation and can cause health problems and disfigurements, sometimes severe and affecting facial aesthetics, with a significant negative impact on a patient's psychology and quality of life. Various skin-whitening agents like hydroquinone, kojic acid, arbutin, azelaic acid, ellagic acid, and resveratrol are used in cosmetic and dermatological products. However, all these agents lack efficacy and safety.[10]

Overall, there is a great need and a massive market for nontoxic skin-whitening agents to fulfil the social needs of safe cosmetic practices and medical hyperpigmentation treatments, and *hs*TYR inhibitors have to be the best choice.

1.3.2 The use of antimelanogenic agents in melanoma treatments

Recently, it was demonstrated that *hs*TYR inhibition could be beneficial in melanoma management.[17] According to the Melanoma Skin Cancer Report (2020),[20] melanoma skin cancer incidence rates rose by 44% between 2008 and 2018, with deaths increasing by 32%. In 2018, 287,723 cases of melanoma skin cancer were diagnosed, and 60,712 people died. This study developed a tool that predicted that by 2025, the number of melanoma skin cancer cases diagnosed worldwide would increase 18% to 340,271, with deaths rising 20% to 72,886. A prevision for 2040 was also made, and an estimated half a million people would be diagnosed with melanoma skin cancer, an increase of 62% on 2018 records, while 105,904 will die from the disease, an increase of 74%.

Melanoma has a high lethality rate due to multiple resistances to current anticancer therapy, particularly in stages III and IV of the disease. At stage IV, the median overall survival with standard dacarbazine chemotherapy is between 6 to 10 months, with a 5-year survival rate of 10%, a value recently extended to 20% using innovative immunotherapy approaches involving specific immune-checkpoints targeting.[21]

Unlike normal human melanocytes, Melanoma cells do not transfer melanin production to neighboring keratinocytes, instead accumulating the pigment. The resulting high concentration of intracellular melanin has several negative consequences. First, melanin has long been recognized for conferring radioprotective effects on melanoma cells, even though radiotherapy, once considered ineffective, is now seen as beneficial in some cases.[22] Recently, a clear relationship between radiotherapy efficacy and melanin accumulation was discovered, implying that inhibiting melanogenesis could sensitize melanoma cells and improve overall outcomes.[10]

However, to the best of our knowledge, no clinical practice or clinical studies, including the use of melanogenesis inhibitors in melanoma therapy, have been documented to date, suggesting an untapped potential. Because the field is still in its early stages, only preliminary results involving melanoma cells in *in vitro* tests and frequently suboptimal *hs*TYR inhibitors like kojic acid are available. Nonetheless, the preliminary findings are quite encouraging and identifying reliable, efficient, and safe *hs*TYR

inhibitors will undoubtedly contribute to the advancement of clinical trials following these pioneer investigations.

1.3.3 Neuromelanin synthesis and Parkinson's Disease

In addition to its role in peripheral (cutaneous and ocular) melanogenesis, *hs*TYR is believed to play an essential role in neuromelanin synthesis in the brain.[10]

Neuromelanin is most pronounced in catecholaminergic neurons of the substantia nigra pars compacta (SNpc) and locus coeruleus, giving these areas of aged brains a blackened appearance. Like the other melanins, neuromelanin presents radical scavenging, antioxidant, toxic metal binding, and toxins sequestering properties to particularly exposed catecholaminergic neurons. However, neuromelanin is also known to be a potential cause of dopaminergic neuron degeneration and, eventually, Parkinson's disease.[10] It is believed that Parkinson's disease is distinguished by the loss of neuromelanin and the subsequent depigmentation of these brain regions.[23] This is primarily due to pigmented cells dying due to immune-mediated death. There is evidence that dopaminergic neurons with high levels of neuromelanin are more prone to degeneration.[24]

While most cancers are less common in Parkinson's disease patients, melanoma is more common than in the general population.[25-27] Bose et al. (2018)[25] reported a large-scale study which found that a melanoma diagnosis is associated with a 50% increased risk of developing Parkinson's disease, and that patients with Parkinson's disease have a 2-fold increased risk of developing melanoma later in life.[25]

Overall, the previously reported aspects point to *hs*TYR playing a significant role in neuromelanin synthesis and in the development of Parkinson's disease. As a result, inhibiting neuronal *hs*TYR function may be a viable exploratory treatment approach for Parkinson's disease. However, it is critical to remember that rigorous regulation of neuromelanin levels between protective and pathogenic thresholds is required and probably challenging.

1.3.4 Tyrosinase and alkaptonuria-related Ochronosisis

Recently, it has been hypothesized that *hs*TYR may have a role in ochronosis that occurs in alkaptonuria. Alkaptonuria is a rare condition caused by an inactivating mutation in the gene that codes for homogentisate 1,2-dioxygenase (HGD), an enzyme

capable of converting homogentisic acid (HGA), a *p*-diphenol intermediate in L-tyrosine metabolism, into linear products via phenyl ring oxidation. While young individuals can remove the amounts of HGA generated daily through urine, the decline in renal function with aging causes HGA buildup in numerous tissues. Ochronosis occurs when non-excreted HGA is oxidized into benzoquinone acetic acid, which is then polymerized into an ochronotic pigment in a very similar process to melanogenesis.[28] The issue associated with the accumulation of this ochronotic pigment is that it causes a dark blue deposition in various tissues like the skin, cartilages, tendons, ligaments, eyes, ears, heart, arterial system, or bones, and its progression is associated with rapid tissue destruction, debilitating clinical morbidities, and eventually death.[28]

Given the enzyme's nonspecific activity, the oxidation of HGA might be mediated by *hs*TYR, opening the way for the emergence of ochronosis. Indeed, TYRs may oxidize *p*-diphenols like hydroquinone under specific circumstances, particularly in a media-rich in catechols like L-DOPA.[9] The participation of *hs*TYR in the disease is not completely established yet. However, if, in the future, *hs*TYR is found to contribute to the spread of the disease, inhibiting the enzyme activity may give a treatment option for avoiding ochronosis in the setting of alkaptonuria.

1.4 Known *hs*TYR inhibitors

In the literature dedicated to TYR inhibitors, an implicit assumption prevails, compounds found using mushroom tyrosinase assays are considered promising TYR inhibitors for human-directed applications. Indeed, tyrosinase from the white button mushroom *Agaricus bisporus* (*ab*TYR) has been widely available from commercial sources for several decades, whereas *hs*TYR is still expensive and difficult to produce.[10] Therefore, the overwhelming majority of TYR-targeting compounds have been identified purely based on anti-*ab*TYR activity. Several thousand *ab*TYR inhibitors are known, including many different scaffolds and a considerable number of natural products. Even the popular kojic acid, commonly utilized in human dermocosmetics, was identified as an *ab*TYR inhibitor in 1979 as part of a phytochemistry research.[26] However, *ab*TYR and *hs*TYR are very different. The *hs*TYR is a highly glycosylated monomeric protein anchored in the melanosome membrane, while *ab*TYR is a soluble oligomeric enzyme found in the cytoplasm.[10]

As shown in **Figure 7**, *hs*TYR has a unique cysteine-rich subdomain located in the cytosol of melanosomes (the EGF domain), a transmembrane hydrophobic domain, and seven asparagine glycosylation sites (N86, N111, N161, N230, N290, N337, N371), three features completely absent from *ab*TYR. Even at the active location, significant changes can be seen. The second coordination sphere varies substantially in the dicopper center surrounded by six histidines. Some hot spot residues around the *hs*TYR active site, such as H304, K306, R308, T343, T352, I368, S375, and S380, that were predicted to have critical direct interactions with some of the most effective *hs*TYR inhibitors discovered to date, are completely missing from *ab*TYR, more properly from the two most abundant *ab*TYR isoforms, being that *ab*PPO3 and *ab*PPO4 (**Figure 7**).[27,29]

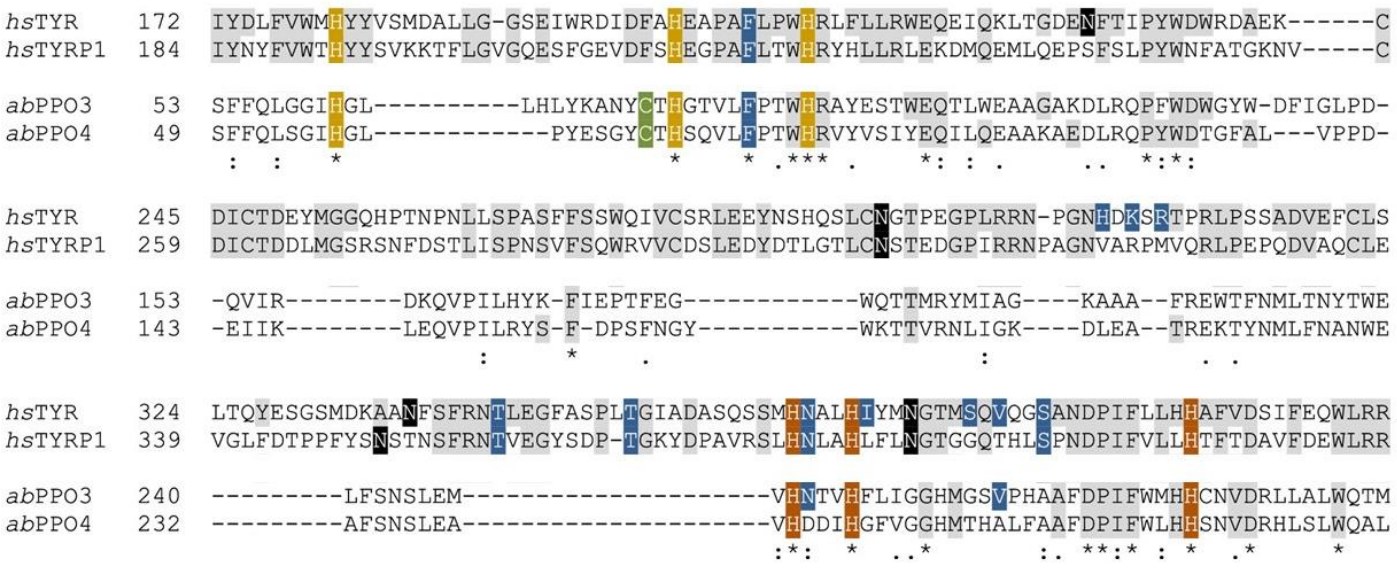


Figure 7 – *hs*TYR, human TY-related protein 1 (*hs*TYRP1), and *Agaricus bisporus* PPO3 (abPPO3) and PPO4 (abPPO4) TY isoforms multiple sequence alignment: Histidine residues bound to CuA/ZnA are shown in yellow; histidine residues bound to CuB/ZnB are shown in orange; glycosylation positions are shown in black; cysteines responsible for the formation of a thioether bond at the active site are shown in green; residues predicted to interact with *hs*TYR inhibitors are shown in blue; *hs*TYR residues conserved in other enzymes are shown in light gray.[10]

As a result, when tested against *hs*TYR, several *ab*TYR inhibitors provide disappointing results, including some of the most extensively used depigmenting compounds now used in dermocosmetics.[10] Because of this general inefficiency, significant concentrations of skin-whitening agents such as kojic acid and hydroquinone are used in dermatological treatments. However, these products carry the risk of inducing side effects such as *contact dermatitis*, an allergic reaction that can cause skin discomfort. With time and long-term use, these whitening agents can make the skin more prone to sunburn.

Several recent studies have concentrated on determining the anti-TYR activity of different compounds employing enzymatic assays and *hs*TYR-overexpressing human cell lines. The assays in human cell lines include melanogenic human malignant melanoma cells (G-361, HBL, HMV-II, SKMEL), normal adult or newborn melanocytes (HEMa and HEMn, respectively), and human nonmelanogenic cells transfected with a *hs*TYR construct (HEK-293-TYR), which readily provide cell-free crude *hs*TYR preparations for inhibitor screening.

In a review by Roulier et al. (2020)[10], a sizable number of natural products and synthetic compounds were tested, and reports describing IC_{50} values below 100 μM were considered. This literature review helps understand which *hs*TYR inhibitors may be more appealing. The authors state that a significant majority of the compounds reported are phenol derivatives. As expected, structural analogues of L-tyrosine and L-DOPA synthesized from p-coumaric acid and caffeic acid exhibited some affinity for *hs*TYR. Flavonoids and derivatives with moderate activity were also found in the flavone dihydrochalcone and aurone subclasses.

Given the nonspecific nature of *hs*TYR catalytic activity, it is expected that at least some of them are substrates rather than inhibitors, resulting in subsequent oxidation of the anti-*hs*TYR agent. When confronted with isolated TYRs from diverse species, agents like resveratrol, all 4-phenols, caffeic acid, and all 3,4-catechols, have demonstrated an alternative substrate behavior, rather than genuinely inhibiting the enzyme function (**Figure 8**, catalysis component). Just like with hydroquinone, the oxidation of depigmenting agents by *hs*TYR might produce reactive quinones and cause the production of potentially toxic polymers or conjugates.

Thus, the authors propose that future research of novel active-site binding *hs*TYR inhibitors should include the use of non-oxidizable equivalents to phenols or catechols, such as resorcinol groups and transition-state analogues. On the other hand, these transition-state analogues are termed by their structural similarity to the catechol substrate and quinone product of the TYR catalytic cycle, but their oxidation state prevents TYR-mediated enzymatic reaction. As a result, when the dicopper core is bound, they exhibit genuine inhibitor behavior (**Figure 8**, catalysis component).

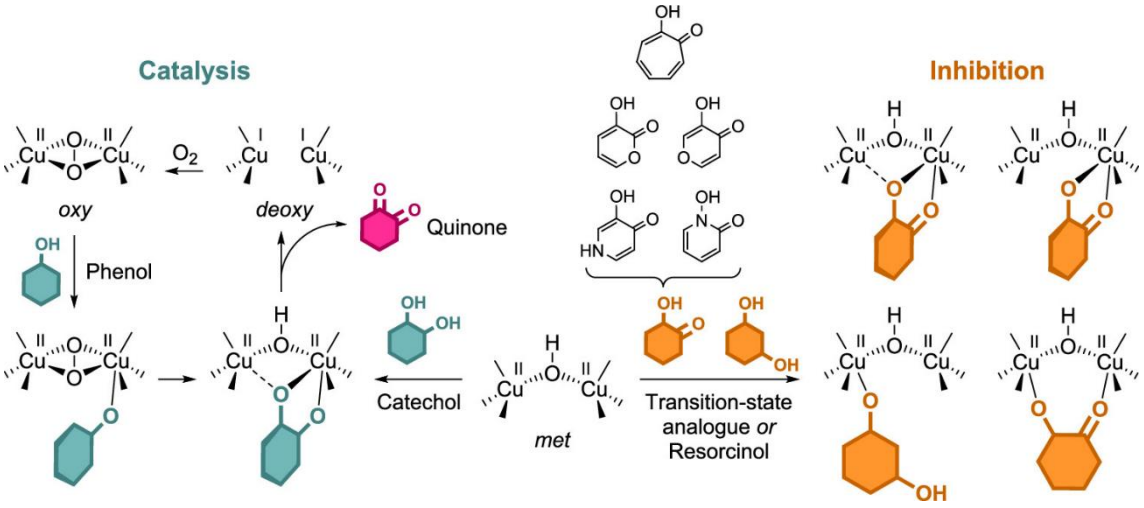


Figure 8 – The catalytic cycle of the oxidation of 4-phenols and 3,4-catechols by TYRs, and examples of potential binding modes for non-oxidizable transition-state analogues and resorcinols as active-site inhibitors.[10]

The discovery of specific *hs*TYR inhibitors can be performed against the isolated enzyme through enzymatic assays. However, only lately, *hs*TYR is being used instead of *ab*TYR, following significant advances in *hs*TYR expression and purification. As it is possible to verify through **Table 1**, from the compounds that are considered to be classical TYR inhibitors, only L-mimosine and phenylthiourea, significantly affected the catalytic activity of the *hs*TYR, while kojic acid, hydroquinone, arbutin and resorcinol displayed almost inactive profiles.[10]

Table 1 – Activity of some inhibitors against Isolated *hs*TYR.[10]

Inhibition value	TYR inhibitor	*hs*TYR activity (µM)	*ab*TYR activity (µM)
K_i	**L-mimosine**	10.3	0.13
	Phenylthiourea	1.7	–
	Kojic acid	350	4.3
	Hydroquinone	4,400	1.1
	Arbutin	6,500	40
	Resorcinol	>3,000	652
	Thiamidol	0.25	–
	Aurone	0.35	–

Presently only two *hs*TYR inhibitors, thiamidol and aurone, were further evaluated in clinical assays.[10] In general, several important conclusions about the development of novel *hs*TYR inhibitors were drawn from the review by Roulier et al. (2020)[10]**:**

1. A single *ab*TYR inhibition measurement should never be used to make a straight assumption about a potential *hs*TYR inhibition activity.

2. A direct measurement of *hs*TYR inhibitory activity against an isolated enzyme structure it's an important step and must be done to obtain valuable kinetic data,

identifying the mechanism of action and providing information about the binding location.

3. Compounds destined for human-directed applications should be evaluated not only against *ab*TYR, but also against cell-free crude extract of a proper *hs*TYR-expressing cell line, such as human melanoma cells, human melanocytes, or transfected nonmelanogenic human cells.

4. In the context of TYR inhibition, the synthesis of phenyl analogues with 4-hydroxy or 3,4-dihydroxy patterns should be avoided since they may operate as alternative substrates rather than inhibitors.

5. In active-site targeting, using transition-state analogues and resorcinols that mirror the structure of naturally oxidized phenolic rings while resistant to *hs*TYR-mediated oxidation has already shown encouraging results.

6. An investment in bioinformatic studies is critical for the development of appropriate methods and tools, particularly for active-site inhibitors, because the description of copper ligand interactions must be accurate enough, as multiple binding modes are frequently conceivable and it partially determines the orientation of the molecule and its ability to reach critical interactions.

1.5 QSAR Modeling

QSAR (quantitative structure-activity relationship) modeling is an *in silico* methodology that aims to predict physical or biological properties of small molecules. In these QSAR studies, a mathematical model is developed that relates the biological activity of the studied compounds to their molecular structure.[30] In general, a QSAR analysis tries to find correlations between a biological activity of interest and molecular descriptors, either calculated from the compound structure or experimentally obtained.[31] QSAR modeling was pioneered by Corwin Hansch 60 years ago and was initially conceptualized as the logical extension of organic chemistry.[32] QSAR modeling has grown, diversified, and evolved from its application to small series of similar compounds, using relatively simple regression methods, to the analysis of much larger datasets spanning thousands of molecules, using a wide variety of statistical and machine learning techniques. Continuous improvements allowed QSAR modeling to be used in chemical, medical and pharmaceutical industries and in government institutions worldwide.

In a first phase, QSAR models attempt to build a link between a collection of chemical structures, their molecular characteristics, and a biological activity. Following that, in a second phase, these models may be used to predict the biological activity of novel chemical compounds having structures comparable to those employed in the QSAR model's construction (**Figure 9**).[33, 34]

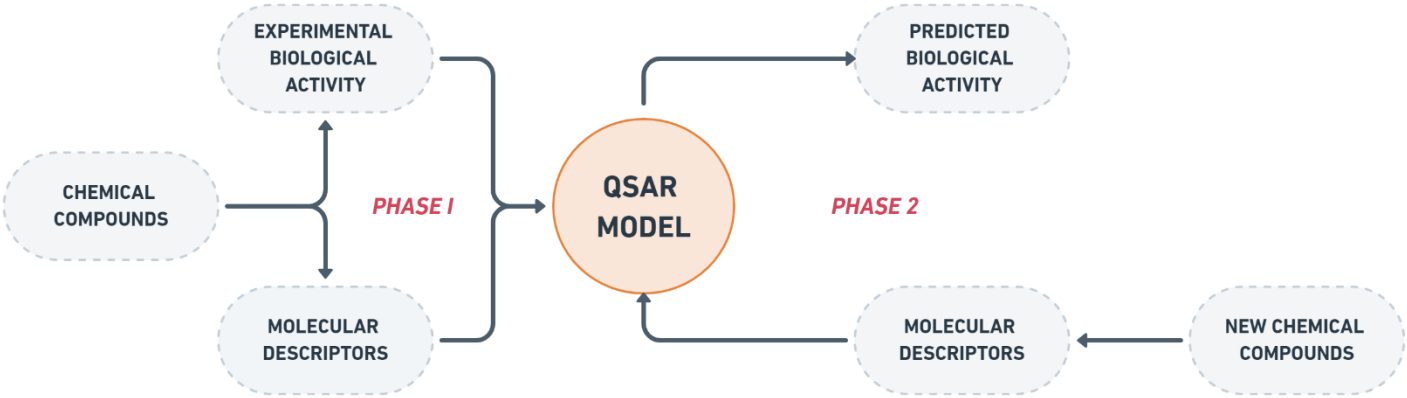


Figure 9 – Representative scheme of the general objective of a QSAR prediction model.

Each substance has several molecular and physical properties known as molecular descriptors. A molecular descriptor results from a mathematical and logical method that converts the information encoded in a molecule *in silico* representation into a useful number. Different molecular descriptors indicate various levels of structural representation. According to Xue L., & Bajorath J. (2000),[35] these descriptors can be categorized based on their "dimensionality" in the following ways:

1. One-dimensional (1D): based on physicochemical properties and the molecular formula (e.g., molecular mass, molar refractivity, logP).
2. Two-dimensional (2D): describe properties that can be calculated from a 2D representation (e.g., number of atoms, number of bonds, connectivity indices).
3. Three-dimensional (3D): depends on the 3D conformation of the molecules (e.g., Van der Waals volume, solvent accessible surface area).

There are other levels of structural representation, such as 4D descriptors, proposed by Hopfinger et al. (1997)[36] that use conformations obtained through molecular dynamics simulation. The 5D descriptors proposed by Vedani A., & Dobler M. (2002)[37] were an extension of the 4D proposed by Hopfinger et al. (1997),[36] adding conformational freedom, thus allowing multiple representations of the ligand topology at

the active site. The same group then proposed 6D descriptors, which consider several solvation models simultaneously.[38]

According to Todeschini et al. (2000),[39] another classification of molecular descriptors concerns their nature, which can be:

1. Constitutional: are derived from the atomic composition of the compound (e.g., molecular weight, number of atoms and bonds).
2. Topological: (e.g., content index of link information).
3. Geometric: are derived from 3D coordinates (e.g., molecular volume, polar surface area).
4. Electrostatics: are derived from partial charges (e.g., polarity indices, partial charges).
5. Quantum mechanics: they are derived from the wave functions of electrons (e.g., energy of molecular orbitals).

Several computer programs are available for calculating molecular descriptors, such as DRAGON®, OCHEM, Mordred, ISIDA, PaDel, among others. When these descriptors can be stated numerically, it is possible to recognize a mathematical link between the descriptors and the biological activity and then obtain a quantitative structure-activity relationship (QSAR) model. In **Figure 10,** an overview scheme with the main QSAR modeling steps is presented.

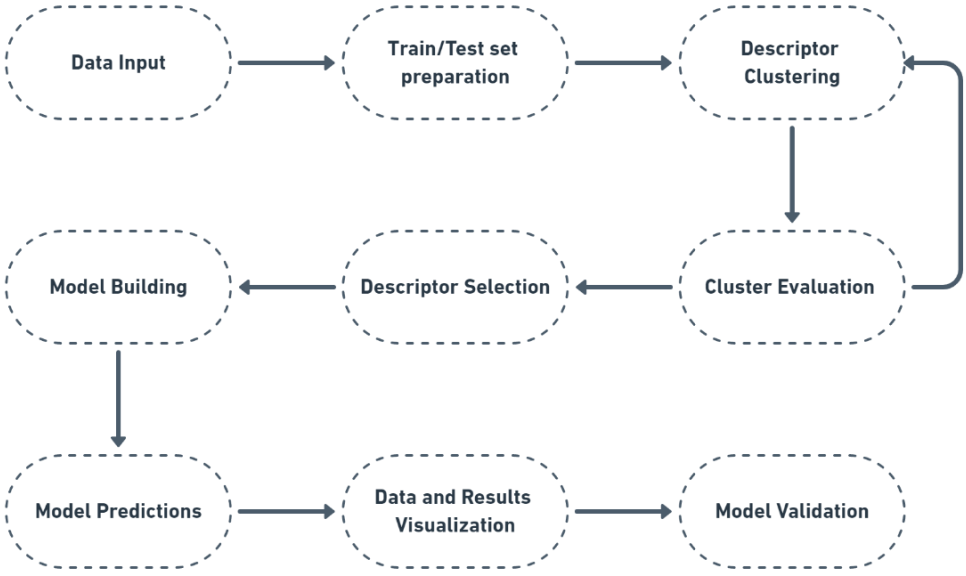


Figure 10 – Scheme demonstrating the fundamental processes for the creation of a QSAR model.

1.6 Compound Library Databases and QSAR Libraries

The creation of a good library of compounds is one of the critical steps to obtain a good QSAR model. For QSAR research, the number of compounds in the dataset should not be too little or, for practical reasons, too big. The top limit is frequently dictated by the computing and temporal resources available for developing QSAR models using the chosen methodologies. For a compound to be part of a library, it must have a numerical value already obtained through experimental tests, and it is possible to find them scattered throughout the literature. For example, in order to carry out this dissertation, a library with 100 compounds previously assembled by Oliveira et al. (2021)[40] will be amplified up to a library with as many compounds available as it is possible to find.

To apply a QSAR model that has already been built and validated, compounds only need to fulfil two requirements, namely:

1. The compounds that are intended to test need to have a structure similar to the compounds used to build the applied QSAR model.
2. Calculate the molecular descriptors that the applied QSAR model indicated as the most relevant for predicting the studied biological activity.

Knowing this, it is possible to easily apply a QSAR prediction model from any database with the structure of the molecule that is wished to employ.

There are many compound databases that provide invaluable help for all types of bioinformatic work. Databases like Drugbank, Pubchem, Chemoinfo and Openmolecules are great examples of suitable free databases. However, among all, the ZINC15 Database stands out for the immense number of compounds it has. This is a free database of commercially available compounds specifically for many types of virtual screening works. ZINC contains over 750 million purchasable compounds and a lot of different mechanisms to facilitate the search for the intended compounds. Some databases only focus on specific characteristics of compounds. A good example of such databases is the COCONUT (COlleCtion of Open Natural prodUcTs). The COCONUT database is a database focused only on natural compounds, which is free and available to all users. **Table 2** presents some general aspects of each Database referred to at this point.

Table 2 – General information about some of the most popular databases

Database	Number of compounds	URL	Description
Drugbank	More than 500 K	https://go.drugbank.com	Database containing information on drugs and drug targets
Pubchem	More than 122 M	https://pubchem.ncbi.nlm.nih.gov	Information on chemical structures, identifiers, biological activities, patents, and many others
Chemoinfo	More than 760 K	https://chemoinfo.ipmc.cnrs.fr/index.html	Downloadable bioinformatics data and tools for small drugs (molecular weight ≤ 2000)
Openmolecules	No information	https://openmolecules.org	Platform to publish cheminformatics tools to contribute for synthetic and medicinal chemistry
ZINC15	More than 750 M	https://zinc15.docking.org	Commercially available compounds for virtual screening
COCONUT	More than 407 K	https://coconut.naturalproducts.net	Natural Products storage, search, and analysis

2. Material and methods

2.1 *hs*TYR Library construction

In order to implement the *in silico hs*TYR library, an extensive literature search was performed. As a selection criterion, the compound had to have been subjected to an *in vitro* enzymatic assay against *hs*TYR, and an IC_{50} value had to be available. All compounds with experimental assays using TYR from non-human species (specifically *ab*TYR) or with *hs*TYR inhibition values obtained by *in silico* studies were disregarded.

A library of 100 compounds was already prepared by Oliveira et al. (2021)[40], and for this work, the objective was to expand the initial library to at least 200 compounds. In the end, 96 compounds were discovered as *hs*TYR inhibitors and added to the library. Therefore, the library prepared and used in this dissertation was composed of 196 compounds.

The *hs*TYR library compounds were divided into groups according to their structural constitution. This division was made according to their structural similarities. Compounds presenting similar scaffolds were grouped, and six groups were defined. **Figure 11** depicts the scaffold templates defined for each group.

The defined groups were as follows:

Group 1 – Compounds formed by a hetero-bicyclic ring system linked to a single aromatic ring (23 compounds).

Group 2 – Compounds formed by one aromatic ring in each terminal linked by a 2-8 carbon linker (32 compounds).

Group 3 – Compounds formed by a hetero-penta ring (39 compounds).

Group 4 – Compounds formed by a tricyclic ring system (41 compounds).

Group 5 – Small compounds formed by a single aromatic ring (32 compounds).

Group 6 – Compounds with at least one glicolisation (33 compounds).

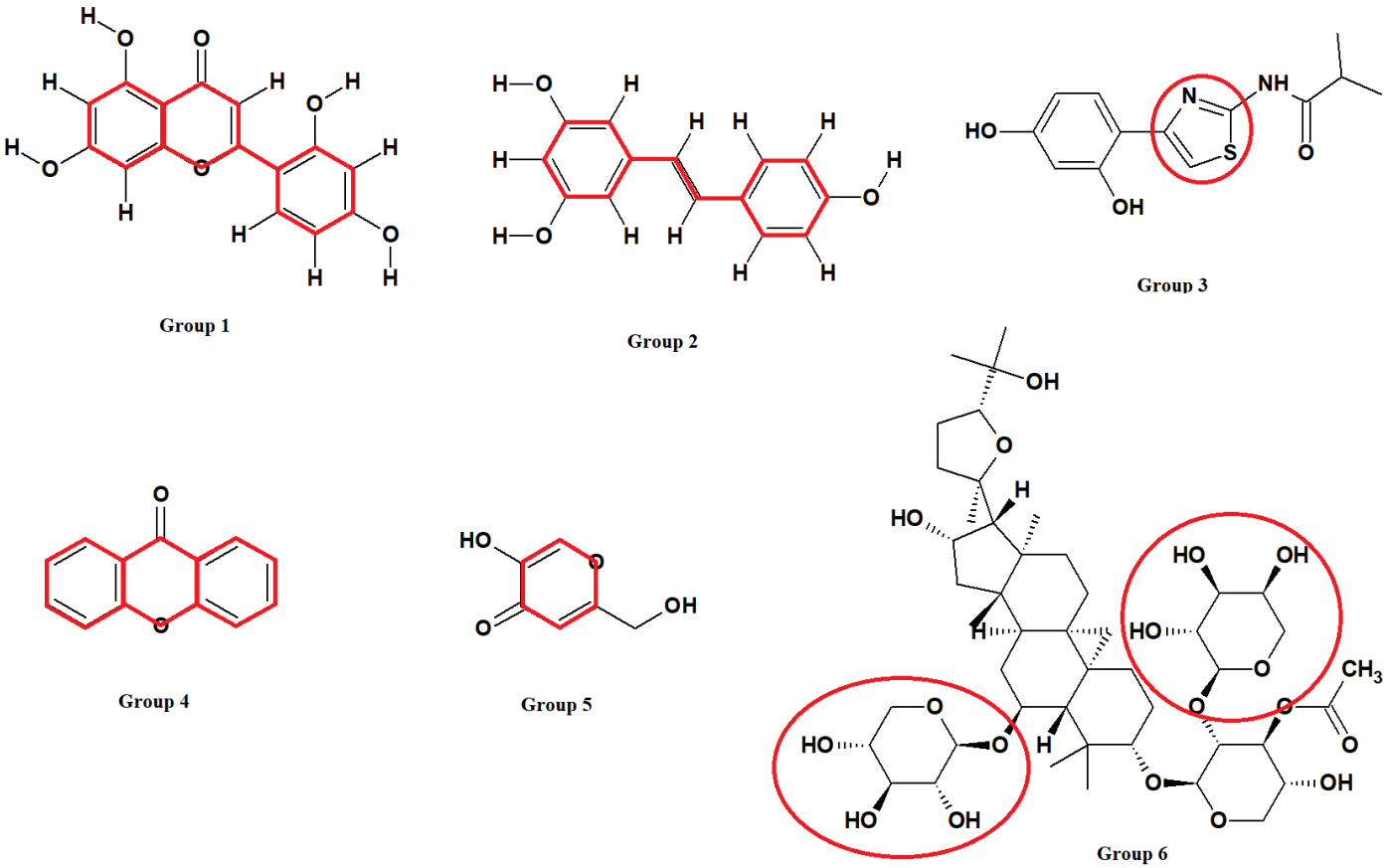


Figure 11 – Example of a representative compound from each group obtained after analyzing the structural similarities of the compounds in the library. The scaffold used to divide the compounds into different groups is represented in red. Group 1 – Norartocarpetin[41], Group 2 – Resveratrol[10], Group 3 – Thiamidol[10], Group 4 – Xanthone[42], Group 5 – Kojic Acid[43], Group 6 – Askendoside B[44]. 2D strucures prepared using ChemSketch software.

For each compound, the experimental IC_{50} values of each was standardized to the same unit of magnitude (Molar) and then converted into experimental pIC_{50} values using the formula (**Table S1**):

$$pIC_{50} = -log\,(IC_{50})$$

2.2 Molecular descriptors calculation

Several benchmarking tests have shown that the choice of statistical techniques has far less of an impact on the prediction performance of QSAR models than the type of molecular descriptors.[45,46] That said, the choice of descriptors is extremely important, which means that the choice of the software or platform for calculating the descriptors is crucial for the success of QSAR modeling. As previously mentioned, several computer programs or platforms are available for calculating molecular descriptors, such as

DRAGON®, OCHEM, Mordred, ISIDA, and PaDel, among others. After analyzing the alternatives, the OCHEM platform was selected. This platform was selected because it is freely available and has proven helpful in other studies to be an effective molecular descriptor calculator.[47]

OCHEM is an online platform that automates and simplifies the typical steps required to build QSAR models. The operation of OCHEM is based on the wiki system and mainly focuses on data quality and verification. The Database that supports OCHEM is fully integrated with the modeling framework, which supports all the steps necessary to create a predictive model, namely:

1. Data search.
2. Selection and calculation of a large variety of molecular descriptors.
3. Application of machine learning methods.
4. Validation of data and QSAR models.
5. Model analysis and applicability domain assessment.

Compared to other similar systems, OCHEM is not intended to re-implement existing tools or models but invites original authors to contribute with their results, making them publicly available and promoting community growth in this area. Iurii Sushko (one of the creators of the online platform) mentioned in the OCHEM user manual: "Our intention is to make OCHEM a widely used platform to perform the QSPR/QSAR studies online and share it with other users on the Web. The ultimate goal of OCHEM is collecting all possible chemoinformatics tools within one simple, reliable and user-friendly resource."[47]

Although OCHEM performs all the tasks described above, it is essential to note that for this work, OCHEM was used only as a tool for calculating the molecular descriptors of the different compounds, and another tool was used to perform the QSAR model.

2.3 PyQSAR and Jupyter Notebook

PyQSAR is a module based on the Python programming language used to develop the QSAR models in this work. PyQSAR seeks to unify, in a single workspace, all the different steps involved in creating a QSAR model, from data preparation to data

validation. Python is one of the most used scripting languages in chemoinformatics and QSAR modeling due to its libraries for numerical analysis, Machine Learning, and plot graphs. Jupyter Notebook it's a software that aims to create a more suitable work environment to produce interactive, appealing, and interesting analyses. This Notebook also has the advantage that the same action can be performed several times and in separate cells. Like most of the QSAR model-building tools, PyQSAR has a similar workflow to that shown in **Figure 10**.

2.3.1 Data Input and Training/Test set preparation

PyQSAR takes as input the experimental values associated with molecules and molecular descriptors generated by a descriptor calculation tool. There are thousands of ways to present the same information depending on the molecular descriptor calculation software chosen, as user preferences for molecular descriptor calculation software vary. This is because not all software will calculate the same number of descriptors or group them in the same way.

In order to combat this diversity of different ways to present molecular descriptors, PyQSAR uses a clustering process, forming groups (Clusters) of descriptors with similar calculation values, always starting from the first descriptor of the input that is provided. The data preparation module in PyQSAR can divide the input data into two sets: the Training Set, used to build the model, and the Test Set, used to validate the model. After this division, only the molecular descriptors associated with the compounds present in the Training Set are clustered, forming groups of highly correlated descriptors to select and discard all information that may be repeated. This procedure prevents two descriptors that qualify the same information from being chosen as a key descriptor for predicting modulated biological activity (**Figure 12**). After forming all groups of molecular descriptors, they will be evaluated and reformed if necessary.

```
In [14]: # clustering
clust = cl.FeatureCluster(X_data, 'average', 3)
clust_info = clust.set_cluster()
```

```
Cluster 2444 ['fliporingC2A..PyDescriptor.']

Cluster 2445 ['fringClipo3A..PyDescriptor.']

Cluster 2446 ['ringC_lipo_1A..PyDescriptor.']

Cluster 2447 ['R7e...alvaDesc.']

Cluster 2448 ['R7i...Dragon7.']

Cluster 2449 ['cddd_91..CDDD.']

Cluster 2450 ['cddd_104..CDDD.']

Cluster 2451 ['PW3..alvaDesc.']

Cluster 2452 ['X2A..Dragon7.']

Cluster 2453 ['Mor08p..alvaDesc.']
```

```
In [15]: clust.cluster_dist()
```

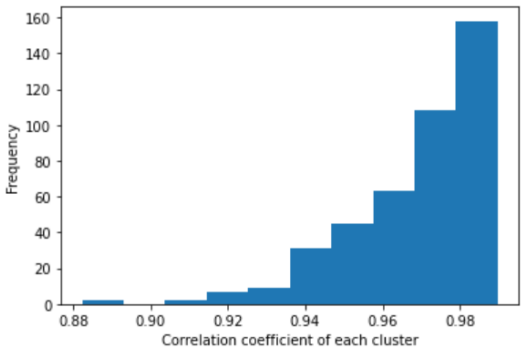


Figure 12 – Print screen taken from Jupyter Notebook with part of the run code to obtain the QSAR model through the PyQSAR module. In the image it is possible to see an example of the clusters formed by the molecular descriptors as well as a graph that demonstrates the frequency of the correlation between the different clusters formed.

2.3.2 Molecular descriptors selection

During the descriptor selection process, a genetic algorithm (GA) was used to maintain the best descriptors from the different Clusters. This method helps to prevent the selection of descriptors with similar properties (belonging to the same cluster). The GA is a method for solving both constrained and unconstrained optimization problems based on natural selection, the process that drives biological evolution. This algorithm repeatedly modifies a population of individual solutions. At each step, the genetic algorithm selects individuals from the current population to be "parents" and uses them to produce the "children" for the next generation. Over successive generations, the population "evolves" toward an optimal solution.

The rationale behind the GA method is basically iterating selections of data until the ideal set of data to obtain the optimal solution to the problem is found. In this case, GA-

based selections were repeated until the optimal set of descriptors was obtained. Each time a new cluster of descriptors was generated during the selection process, a multiple linear regression (MLR) was performed, with the set of descriptors generated to obtain the coefficients of a given set of descriptors. MLR is a statistical method for predicting the result of a response variable by using several explanatory factors. Modeling the linear connection between the explanatory (independent) factors and response (dependent) variables is the aim of multiple linear regression. The *hs*TYR inhibiting power of the compounds, the value predicted with the QSAR model, was used as an independent variable in the MLR process. The MLR score is expressed as a determination coefficient (R^2) and has a value between 0 and 1. The closer this score is to 1, the better its predictive capacity of the QSAR model.

After testing hundreds or even thousands of possible descriptor combinations, the descriptor sets were organized according to their score obtained from the MLR method, and only a predefined number of them advanced to the next step of the iterative selection process. This final step, for the construction of the model itself, serves to choose the group of descriptors with the best score, and although PyQSAR applies an MLR to all groups and tests all possible solutions, only the best solution found is presented.

2.3.3 QSAR model validation and visualization

After the GA has clustered the different descriptors, these groups are submitted to the MLR process, and the best set of descriptors given the characteristics of the provided molecules is obtained. PyQSAR uses these descriptors' values to predict the biological activity studied and has several features for validating models and viewing the results. One such validation method is cross-validation k-fold.

Cross-validation is a statistical method used to gauge the expertise of machine learning models such as QSAR models based in GA. Because it is simple to comprehend, implement and produce skill estimates that often have a more negligible bias than other approaches, it is frequently used in applied machine learning to compare and select a model for a specific predictive modeling issue.[48] In short, this method divides the total data set into "K" subsets of the same size and from there, one of the subsets is used for testing while the remaining "K-1" is used to estimate the statistical parameters. This validation checks the model's accuracy in predicting the biological activity of novel substances and offers crucial statistical evidence supporting that claim (**Figure 13**). In

this figure it is possible to see an example of the statistical values and the graph that combines the experimental pIC_{50} values with the pIC_{50} values predicted by the model.

```
In [25]: from pyqsar import cross_validation as cv
         cv.k_fold(X_data, y_data, feature_set, k=5, run=100)
```

```
R^2CV mean: 0.91457
Q^2CV mean: 0.892212
RMSE CV : 0.187834
Features set = ['C.026..Dragon7.', 'DISSM2C..Mera.', 'MaxdssC..alvaDesc.', 'WHALES90_Rem..alvaDesc.']
Model coeff =  [[ 1.80793236 -0.87699183  0.87830353 -1.3406125 ]]
Model intercept =  [4.39562116]
```

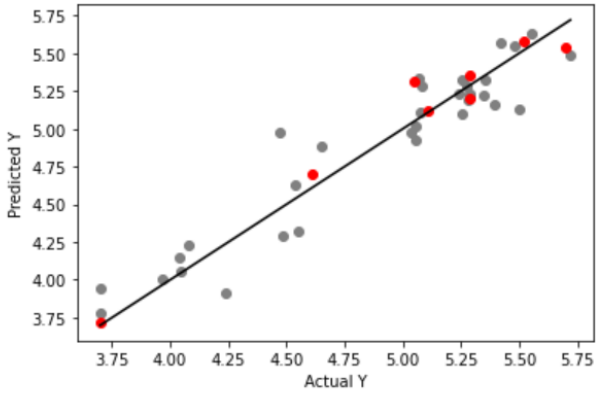


Figure 13 – Print screen taken from Jupyter Notebook with part of the run code to obtain the QSAR model through the PyQSAR module. The red dots correspond to points that were adjusted after using the Cross-Validation method of validation.

2.4 Preparation of the test library and application of the model

After the elaboration and validation of the QSAR model, a library of compounds to be tested must be developed. The following rules must regulate this library:

1. The constituent compounds should have a structure similar to those used to build the QSAR model.
2. It cannot contain repeating molecules.
3. Compounds included in the library must be of natural nature.

Intensive research must be done to gather compounds to integrate the library. To help with this research, several databases provide essential help. Compound databases are abundant and offer tremendous assistance for all bioinformatic tasks. Good free databases include those like Drugbank, Pubchem, ZINC15, EBI, chemoinfo, and openmolecules. The ZINC15 Database distinguishes out from the rest due to the enormous amount of chemicals it contains.

Initiated to provide easy access to compounds for virtual screening, ZINC15 is a public access database and tool set that is now often used for virtual screening, ligand discovery, pharamcophore screens, benchmarking, and force field building.[49] This platform contains a brand-new research method for ligand discovery that links biological actions of medicines, natural chemicals, and gene products with marketability. ZINC15 has over 750 million buyable compounds in its Database, making it quick and easy to find analogs. Additionally, just in ready-to-dock, 3D forms, this widely used Database comprises more than 230 million buyable substances.

For this work, it was defined the focus in the natural component of the compounds gathered, finding the most compounds that might fit into a library to be evaluated and test the developed QSAR model. The ZINC15 Database's catalog of natural products contains over 200 000 items, some of which may be purchased, but no more details are given about them other than their structure and the fact that they are natural products. After researching several databases, the COCONUT database stood out. Its online interface enables a variety of basic searches, including those using molecule names, InChI keys, SMILES, and drawn structures, as well as sophisticated searches using molecular characteristics, substructure searches, and similarity searches.[50] Additionally, users may download the entire dataset or the search results in several formats. The web interface, the back-end, and the Database are all hosted as Docker containers, making it simple to move and deploy on local installations as well as host additional collections of natural products. COCONUT data is extracted from 53 various data sources and several manually collected from literature sets.[50] There are 406,747 distinct "flat" (without stereochemistry) natural products in the most recent COCONUT release, and there are a total of 731,112 natural products whose stereochemistry has been retained the last time this Database was accessed by the research team.[51]

With the library of compounds to be tested complete, it is possible to proceed to the application of the obtained QSAR model. It only takes 3 steps to apply a QSAR model, namely:

1. Calculate molecular descriptors for the compounds to be tested.
2. Obtain the prediction formula of the studied biological activity.
3. Apply the values of the molecular descriptors in the respective formula.

In order to carry out the first step, the same molecular descriptor tool that is used to calculate the molecular descriptors is utilized as variables in the construction of the QSAR model. After the calculation of these molecular descriptors, the calculation formula for the prediction of the studied biological activity follows. This step is relatively simple as PyQSAR provides all the data needed to obtain the formula. This formula can be written as:

$$\boldsymbol{pIC_{50\ PREDICTED} = Intercept + (D_n \times DC_n)}$$

Where $\mathbf{pIC_{50\ PREDICTED}}$ value corresponds to the value of the *hs*TYR inhibition activity predicted by the model, **Intercept** represents the mean value of the response variable when all the predictor variables in the model are equal to zero, $\mathbf{D_n}$ corresponds to the fixed value obtained by PyQSAR for each descriptor that the algorithm defines as the final descriptor that integrates the equation, with $\mathbf{n}$ being replaced by the number of descriptors defined by PyQSAR. $\mathbf{DC_n}$ corresponds to the descriptor value calculated by the molecular descriptor calculation software. Using the same calculation tool that was used to calculate the molecular descriptors employed in the model, the values of the descriptors defined by PyQSAR must be calculated for each new compound that is intended to predict the biological activity studied. The $\mathbf{n}$ is replaced by the number of descriptors defined by PyQSAR.

3. Results and Discussion

3.1 QSAR model selection

For this dissertation, 33 QSAR models were made using the library of compounds gathered and explained in section 2.1. These models were prepared considering the different chemical structure scaffolds in the library of *hs*TYR inhibitors. As stated in the material and methods section, the complete library was divided into 6 groups (G1-G6), and the division was made according to the structural similarities of the compounds (Figure 11). On this stage a decision was made to not consider G6 compounds due to the complexity of their structures. These compounds present a variable number of glycosylation, producing a diversity of chemical structures, both in size and chemical characteristics. This diversity is usually to conducive to good QSAR models.

This separation was performed as it is widely acknowledged that QSAR modelling usually yields better results when the compounds used are similar in structure. For this dissertation, the various groups obtained by splitting the compound library were tested separately, and 6 models were prepared (Table 3). Other models were also obtained by joining groups with structural similarities in their scaffolds. For example, G5 pyrone scaffold compounds were tested with G1 and G3, which also present similar rings in their structure in total. From these combinations, 5 other models were prepared.

Finally, all the 11 models described above were prepared using different sets of molecular descriptors present in OCHEM. For example, the models were prepared using only PADEL descriptors (total of 17,967 descriptors), PADEL descriptors with no fingerprint (total of 1,875 descriptors) and with all the OCHEM molecular descriptors (total of 28,933 descriptors). This procedure has tripled the number of QSAR models built to 33.

The statistical data resulting from the models performed are shown in Table 3. Several models presented good statistical validation, with the top-ranked models being models 23, 29, 32 and 33. These models were all prepared using the complete set of OCHEM molecular descriptors. When analyzed in more detail, it was observed that all these models were obtained with no combinations. Model 23 was obtained from G1 (23 compounds), model 29 was obtained from G2 (32 compounds), model 32 was obtained from G4 (41 compounds), and model 33 was obtained from G5 (28 compounds). All

models obtained from combinations of groups presented worse statistical results. This result is in accordance with the state-of-the-art QSAR modelling knowledge, as it is widely known that QSAR models tend to present a better statistical profile when the library of compounds is co-similar in structure. Libraries with compounds presenting very different structure profiles tend to present the worst predictive capability.

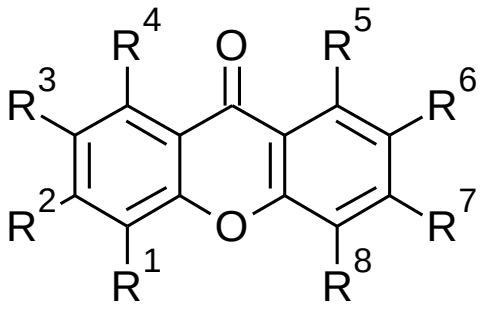
Although all four highlighted QSAR models could be used for further analysis, as they present good statistical validation for all parameters, the decision was made to use model 32 for the remaining of this work. This decision was based on the fact that model 32 was prepared using G4 data matrix, which presents the highest number of compounds (41). The number of compounds used for modelling is essential as it is good practice to use the highest number of compounds to increase the QSAR model's predictive power and statistical robustness. In addition to the greater number of compounds used, this model also presented the lowest values for the errors associated with the model, namely, RMSE=0.1708 and RMSE CV=0.1878. However, models 23, 29 and 33 will also be considered for further use in the subsequent studies.

Table 3 – Statistical results of the 33 QSAR models performed. Statistical data of QSAR model 32 is highlighted. R^2 – determination coefficient. RMSE – square root of the mean error. R^2CV – determination coefficient after use of the cross-validation method. RMSE CV – square root of the mean error after use of the cross-validation method. Q^2CV – average value of the MLR method score five times the cross-validation was performed.

MODEL DESCRIPTION	MODEL NUMBER	R^2	RMSE	R^2CV	Q^2CV	RMSE CV	NUMBER OF COMPOUNDS	GROUPS USED
PADEL DESCRIPTORS	**1**	0.9134	0.3301	0.9167	0.8568	0.3989	23	G1
	2	0.6526	0.5833	0.5853	0.5637	0.6675	94	G1+G2+G3
	3	0.4325	0.7043	0.4370	0.3819	0.7313	163	G1+G2+G3+G4+G5
	4	0.5041	0.7000	0.5450	0.4848	0.7088	124	G1+G2+G4+G5
	5	0.7873	0.3850	0.7916	0.7462	0.4197	64	G1+G4
	6	0.6896	0.5132	0.6927	0.6624	0.5374	92	G1+G4+G5
	7	0.9119	0.3494	0.9146	0.8752	0.4026	32	G2
	8	0.6123	0.6103	0.6160	0.5808	0.6354	99	G2+G3+G5
	9	0.8704	0.2556	0.8723	0.6902	0.4303	39	G3
	10	0.8990	0.1841	0.8999	0.8813	0.1946	41	G4
	11	0.9337	0.2770	0.9236	0.8940	0.2936	28	G5
PADEL DESCRIPTORS NO FINGERPRINTS	**12**	0.8719	0.4016	0.8734	0.8123	0.4483	23	G1
	13	0.4215	0.7689	0.4247	0.3828	0.7904	94	G1+G2+G3
	14	0.3147	0.7739	0.3186	0.2843	0.7923	163	G1+G2+G3+G4+G5
	15	0.3840	0.7801	0.3873	0.3417	0.7997	124	G1+G2+G4+G5
	16	0.7143	0.4463	0.7217	0.6475	0.4989	64	G1+G4
	17	0.5874	0.5925	0.5971	0.5709	0.6145	92	G1+G4+G5
	18	0.7875	0.5426	0.7962	0.7098	0.6451	32	G2
	19	0.5016	0.6920	0.5052	0.4875	0.7092	99	G2+G3+G5
	20	0.8147	0.3057	0.8157	0.6160	0.4882	39	G3
	21	0.8589	0.2175	0.8610	0.8326	0.2339	41	G4
	22	0.9150	0.2924	0.9176	0.8665	0.3471	28	G5
OCHEM DESCRIPTORS	**23**	0.9729	0.1848	0.9732	0.9561	0.2100	23	G1
	24	0.5486	0.6792	0.5525	0.5031	0.7074	94	G1+G2+G3
	25	0.4186	0.7129	0.4221	0.3871	0.7321	163	G1+G2+G3+G4+G5
	26	0.4948	0.7007	0.4971	0.4359	0.7309	124	G1+G2+G4+G5
	27	0.8367	0.3374	0.8400	0.7524	0.4417	64	G1+G4
	28	0.6809	0.5203	0.6851	0.6197	0.5812	92	G1+G4+G5
	29	0.9299	0.3117	0.9318	0.9074	0.3466	32	G2
	30	0.5544	0.6544	0.5585	0.5194	0.6821	99	G2+G3+G5
	31	0.8932	0.2321	0.8951	0.8592	0.2574	39	G3
	32	0.9128	0.1708	0.9147	0.8922	0.1878	41	G4
	33	0.9359	0.2540	0.9379	0.9077	0.2981	28	G5

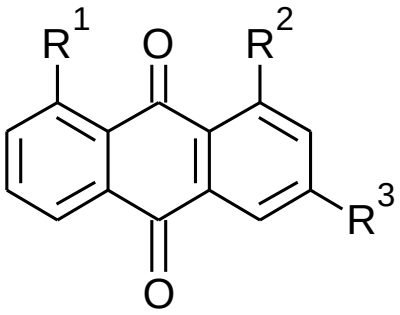
3.2 Structural analysis of G4 compound library

As stated before, QSAR model 32 was selected for further studies. When analyzed in more detail, this group is formed by a total of 41 compounds with a scaffold presenting a tricyclic ring system (**Figure 11**). All compounds are represented in **Figure 14**, **Figure 15**, **Figure 16**.



COMPOUND NUMBER	R^1	R^2	R^3	R^4	R^5	R^6	R^7	R^8	IC_{50} EXPERIMENTAL (μM)
1[42]	-	-	-	-	CH_3	-	OCH_3	OCH_3	2.000
2[42]	-	-	-	-	CH_2O	-	OCH_3	OCH_3	29.130
3[42]	-	-	-	-	-	-	OCH_3	OCH_3	5.120
4[42]	-	-	OH	-	OH	-	-	-	4.520
5[42]	-	OH	-	-	CH_2O	-	OCH_3	OCH_3	3.160
6[42]	-	-	-	-	CH_3	-	OH	OH	8.930
7[42]	-	-	-	-	-	-	OH	OCH_3	3.810
8[42]	-	-	-	-	-	OH	-	-	5.600
9[42]	-	-	-	-	OH	OH	-	-	7.800
10[42]	-	-	-	-	OCH_3	OCH_3	-	-	4.020
11[42]	-	OCH_3	-	-	CH_3	-	OCH_3	OCH_3	22.400
12[42]	-	-	-	-	CBr_2	-	OCH_3	OCH_3	33.700
13[42]	-	-	-	-	-	OCH_3	OCH_3	-	8.290
14[42]	-	-	-	-	-	-	OH	-	9.260
15[42]	-	-	-	-	OCH_3	-	-	-	5.140
16[42]	-	-	-	-	-	-	OCH_3	OH	5.200
17[42]	-	OH	-	-	-	-	OH	-	4.420
18[42]	-	-	-	-	OH	OCH_3	OH	-	3.010
19[42]	-	OH	-	-	CH_3	-	OH	OH	3.280
20[42]	-	-	-	-	OH	-	-	-	8.830
21[42]	-	-	-	-	-	-	OCH_3	-	5.140
22[42]	-	-	-	-	-	-	-	OH	8.460
23[42]	-	OCH_3	-	-	CH_3	Br	OCH_3	OCH_3	5.570
24[42]	-	-	-	-	-	-	OH	OH	5.120
25[42]	-	OCH_3	-	-	CH_3	Cl	OCH_3	OCH_3	1.900
26[42]	-	-	-	-	-	OH	OH	-	5.700
27[42]	-	-	-	-	-	-	-	OCH_3	5.330
28[42]	-	-	-	-	-	-	-	-	8.840
29[42]	CH_2O	-	OH	OH	-	-	-	-	89.370
30[42]	-	-	-	-	-	OCH_3	-	-	8.470

Figure 14 – Scaffold structure for some compounds belonging to G4 compound library and representative table of the R groups present in these compounds.



COMPOUND NUMBER	R^1	R^2	R^3	IC_{50} EXPERIMENTAL (µM)
31[53]	OCH_3	OCH_3	CHN-Thiourea	28.050
32[53]	OH	OH	OCH_3	108.620
33[53]	OCH_3	OCH_3	CH_2OH	200.000
34[53]	OH	OH	$CHNOC_2H_5$	200.000
35[53]	OC_3H_5	OC_3H_5	CHNOH	83.170
36[53]	OH	OH	CHN-Thiourea	24.520
37[53]	OC_4H_9	OC_4H_9	CH_2OH	32.810
38[53]	OC_2H_3	OC_2H_3	CH_2OH	90.990
39[53]	$O_2C_2H_5$	$O_2C_2H_5$	CH_2OH	200.000
40[53]	$O_2C_2H_5$	$O_2C_2H_5$	CHN-Thiourea	58.060

Figure 15 – Scaffold structure for some compounds belonging to G4 compound library and representative table of the R groups present in these compounds.

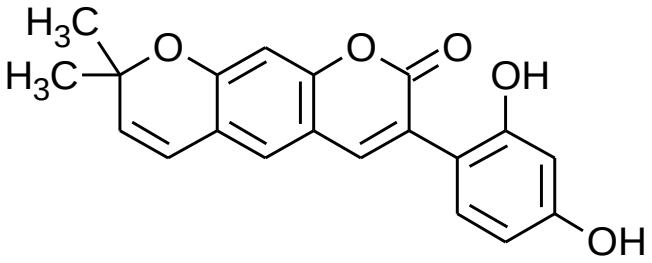


Figure 16 – Compound number 41[54] belonging to G4 compound library with an IC_{50} EXPERIMENTAL of 2.800 µM.

3.3 Detailed analysis of QSAR model 32

Analyzing and comprehending the various statistical data obtained by performing a QSAR model is essential. This analysis will help determine the predictive power of the QSAR model concerning the biological activity, in this case, *hs*Tyr inhibition activity. PyQSAR calculates many statistical data that may be used to determine whether the QSAR model is reliable. Also, the analysis of the molecular descriptors chosen by PyQSAR to construct the final QSAR model is important, as it may provide insights into the favourable characteristics of the compounds used. In this section, the process of

molecular descriptors selection, possible descriptor significance and the statistical parameters calculated will be discussed.

As previously mentioned, this dissertation used the OCHEM platform for molecular descriptors calculation. OCHEM calculated a total of 28,933 molecular descriptors for each computer. However, due to computer processing difficulties, building QSAR models with this number of compounds was time-consuming and prone to errors. So, in order to build the QSAR models, a statistical treatment was performed on the total number of molecular descriptors obtained. An analysis of the variance between the different molecular descriptors was performed using RStudio software, thus reducing the number of descriptors. This analysis grouped descriptors that presented a determination coefficient greater than 0.99. Only one molecular descriptor of each group was considered, and the rest were not considered. By applying this treatment, the number of highly collinear descriptors was reduced and, in the end, from the initial 28,933 molecular descriptors submitted to the treatment, 3,126 remained. These molecular descriptors were submitted to treatments performed by PyQSAR and described in point **2.3.2** of this dissertation, and, in the end, four descriptors were selected as the final descriptors for this QSAR model. These descriptors were:

1. C-026:(Dragon7) – identification of a structural segment, in which case it identifies the structural sequence R--CX--R where: R represents any group linked through carbon, X represents any electronegative atom (O, N, S, P, Se, halogens) and the two hyphens (--) represents an aromatic bond as in benzene.[55]

2. DISSM2C:(Mera) – dissymmetry about the second principal rotational invariant.[56]

3. MaxdssC:(alvaDesc) – Maximum atom-type E-State in the sequence =C< where equal sign (=) represents a double bond, C represents a carbon atom and less than sign (<) represents two bonds to the carbon atom.[57]

4. WHALES90_Rem:(alvaDesc) – WHALES (Weighted Holistic Atom Localization and Entity Shape) Remoteness (percentile 90).[58]

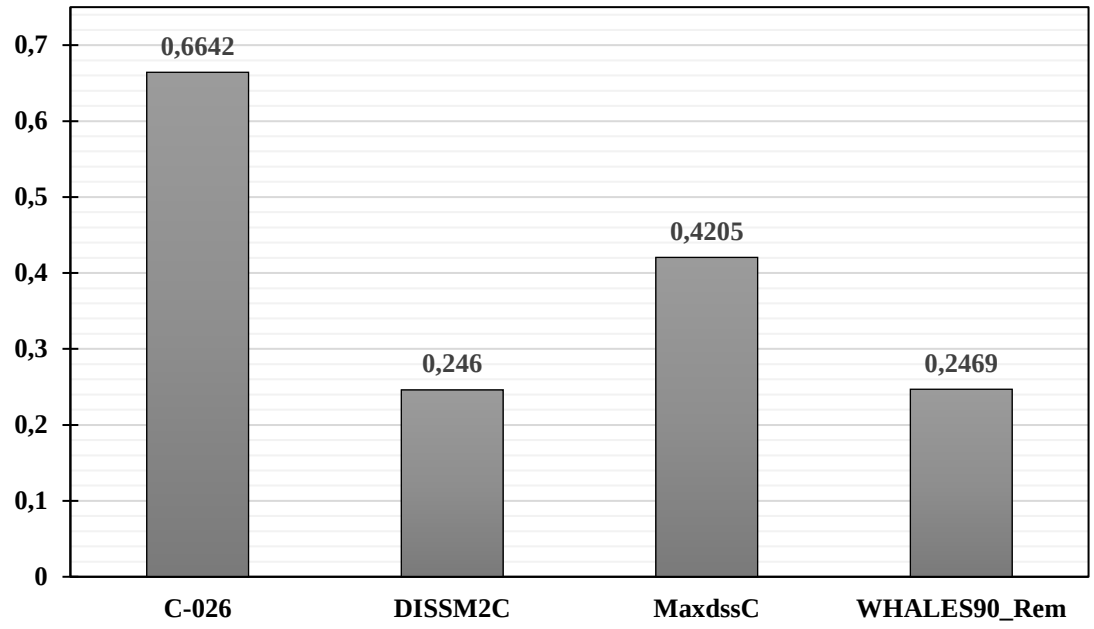
After choosing these molecular descriptors as the most influential for the relationship between G4 compounds structure and their ability to inhibit *hs*TYR, PyQSAR provides

statistical and graphical information that is useful to understand the reason for this choice. One of this information is the correlation matrix between the chosen molecular descriptors. This matrix is presented in **Table 4**, and it is possible to confirm that the molecular descriptors do not have a strong correlation with each other. This information is important as it shows that each molecular descriptor provides complementary information to the QSAR model 32. Each molecular descriptor describes different characteristics of the molecules, which means that a low correlation between them is ideal.

Table 4 – Correlation values between the molecular descriptors.

	C-026	DISSM2C	MAXDSSC	WHALES90_Rem
C-026	1.0000	0.0355	0.0065	0.3656
DISSM2C	–	1.0000	0.2535	0.0571
MAXDSSC	–	–	1.0000	0.0765
WHALES90_Rem	–	–	–	1.0000

Another helpful information presented by the PYTHON language package is the relationship of *hs*TYR inhibition activity, $pIC_{50\ EXPERIMENTAL}$, and the values of the chosen molecular descriptors. This relationship can be seen in **Graph 1** where it is possible to realize that all molecular descriptors contribute to predicting the values of the *hs*Tyr inhibition.



Graph 1 – Correlation values between the molecular descriptors and $pIC_{50\ EXPERIMENTAL}$.

Analyzing **Graph 1,** it is clear that some descriptors contribute more than others because their relationship with the $pIC_{50\ EXPERIMENTAL}$ value is higher. However, it is visible that C-026 descriptor is the one that has a better correlation. Analyzing the significance of each molecular descriptor is sometimes difficult, as they usually are calculated by equations that are not always easy to understand. Also, sometimes there is no information on how the descriptors are calculated.

Of the four calculated descriptors, C-026 is the easiest to understand. It states that compounds with features similar to **Figure 17A** scaffold are favored. An example of a compound with these features is compound 25 (**Figure 17B**), which presents the highest inhibition power of compounds from G4, with an IC_{50} value of 1.9 μM (**Figure 14**). This compound presents the favoured features by C-026 descriptor, with a carbonyl (C=O) group as the CX group, functioning as a linker between two benzene rings. The other descriptors are more challenging to understand, and no helpful information could be extracted concerning favoured features.

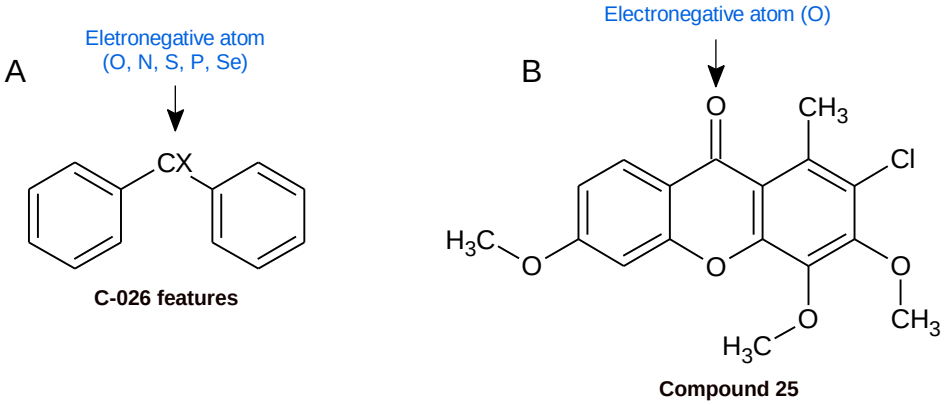


Figure 17 – In **A** it is possible to observe a basic structure of how the descriptor C-026 can be obtained. In **B**, compound 25 can be seen, an example of a structure with the characteristics of the descriptor C-026 present in the library of compounds.

The QSAR model 32 equation developed by PyQSAR, using the selected descriptors, is represented below (and in section **2.4** of this dissertation). Also, the most important statistical parameters are shown and will be discussed. This final equation will later be used to predict the *hs*TYR inhibition capacity of other compounds structurally similar to those used in constructing the QSAR model.

$$\mathbf{pIC_{50\ PREDICTED}} = 4.3956 + 1.8079 \times \text{“}\mathbf{C\text{-}26}\text{”} + (-0.8770) \times \text{"}\mathbf{DISSM2C}\text{"} + 0.8783 \times \text{"}\mathbf{MaxdssC}\text{"} + (-1.3406) \times \text{"}\mathbf{WHALES90\_Rem}\text{"}$$

- **N**=41; **R^2**=0.9128; **RMSE**=0.1708;
- **R^2CV**=0.9147; **Q^2CV**=0.8922; **RMSE CV**=0.1878;

Were **N** corresponds to the number of compounds used; **R^2** is the determination coefficient found when using the Linear Regression method between the various points obtained by PyQSAR when performing the prediction of the biological activity studied for each compound used to build the QSAR model; **RMSE** is the square root of the mean error; **R^2CV** and **RMSE CV** are the values corresponding to **R^2** and **RMSE** after performing the Cross-Validation validation method. Finally, **Q^2CV** is the average value of the MLR method score five times the Cross-Validation was performed.

These statistical parameters generally validate the QSAR model's predictive power. The robustness of the model is essential for its use in accurately predicting the *hs*TYR inhibition activity of other compounds. For this type of QSAR model, there is no statistical reference or threshold beyond which a QSAR model is considered to have good predictive capacity. However, several QSAR models published in the literature consider that a QSAR model has good predictive power if the determination coefficient value R^2 is higher than 0.750 and the root mean error square RMSE value is less than 0.300. All these statistical parameters provide helpful information about the reliability of the model obtained; however, the most important parameter to understand the model's good predictive capacity is Q^2CV. This value corresponds to the average value of the MLR method score after the Cross-Validation is performed five times. This means that this value defines the ability of the model to predict values close to the experimental values; that is, this value defines the model's accuracy. If the Q^2CV value is greater than 0.750, it is usually assumed that the model has good accuracy.

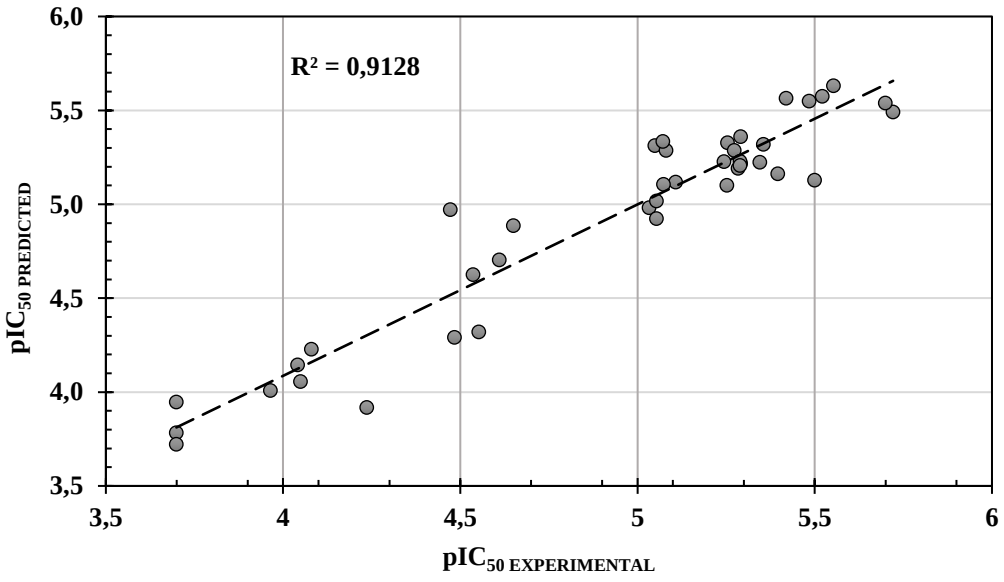
For QSAR model 32, all statistical parameters fall within the values indicated with **R^2** (0.9128); **R^2CV** (0.9147) and **Q^2CV** (0.8922) presenting values well above 0.750; and **RMSE** (0.1708) and **RMSE CV** (0.1878) values falling well below 0.3.[59]

Table 5 presents the difference between the experimental and predicted pIC_{50} values of *hs*TYR inhibition for each of the 41 compounds from G4, used to implement QSAR model 32. The difference is presented as module values, and in general, the differences observed for most compounds are low, ranging from 0.006 (compound 15) to 0.500 (compound 22).

Table 5 – Experimental and predicted pIC_{50} using QSAR model 32 for G4 compounds.

COMPOUND NUMBER	pIC_{50} EXPERIMENTAL	pIC_{50} PREDICTED	$\lvert\Delta pIC_{50}\rvert$	COMPOUND NUMBER	pIC_{50} EXPERIMENTAL	pIC_{50} PREDICTED	$\lvert\Delta pIC_{50}\rvert$
1[42]	5.699	5.539	0.160	**22**[42]	5.073	5.105	0.032
2[42]	4.536	4.625	0.089	**23**[42]	5.254	5.328	0.074
3[42]	5.291	5.216	0.075	**24**[42]	5.291	5.359	0.068
4[42]	5.345	5.224	0.121	**25**[42]	5.721	5.491	0.230
5[42]	5.500	5.128	0.372	**26**[42]	5.244	5.227	0.017
6[42]	5.049	5.313	0.264	**27**[42]	5.273	5.286	0.013
7[42]	5.419	5.565	0.146	**28**[42]	5.054	5.018	0.036
8[42]	5.252	5.101	0.151	**29**[42]	4.049	4.055	0.006
9[42]	5.108	5.118	0.010	**30**[42]	5.072	5.333	0.261
10[42]	5.396	5.161	0.235	**31**[53]	4.552	4.319	0.233
11[42]	4.650	4.885	0.235	**32**[53]	3.964	4.009	0.045
12[42]	4.472	4.972	0.500	**33**[53]	3.699	3.721	0.022
13[42]	5.081	5.287	0.206	**34**[53]	3.699	3.946	0.247
14[42]	5.033	4.981	0.052	**35**[53]	4.080	4.228	0.148
15[42]	5.289	5.228	0.061	**36**[53]	4.610	4.704	0.094
16[42]	5.284	5.191	0.093	**37**[53]	4.484	4.291	0.193
17[42]	5.355	5.319	0.036	**38**[53]	4.041	4.144	0.103
18[42]	5.521	5.575	0.054	**39**[53]	3.699	3.783	0.084
19[42]	5.484	5.548	0.064	**40**[53]	4.236	3.917	0.319
20[42]	5.054	4.924	0.130	**41**[54]	5.553	5.631	0.078
21[42]	5.289	5.206	0.083	–	–	–	–

Graph 2 presents the graphical relationship between pIC_{50} experimental values and pIC_{50} predicted values obtained. The linear regression line that defines the Determination coefficient (R^2) of QSAR model 32 is also presented.



Graph 2 – Graphical representation of the relationship between $pIC_{50\ EXPERIMENTAL}$ and $pIC_{50\ PREDICTED}$

In general, by observing the dispersion of points shown in **Graph 2**, and the statistical values presented, it is possible to infer that QSAR model 32 will potentially provide good predictions of *hs*TYR inhibitory activity when applied to other compounds or interest.

3.4 QSAR model 32 application using natural compounds

Upon validation, QSAR model 32 was used to predict the *hs*TYR inhibition activity of a library of selected natural compounds. This library of compounds was selected using the COCONUT platform with the help of scaffold search tools present in this Database.

The COCONUT database presents several tools to search for structures similar to a given scaffold. For example, it is possible to search for structure similarity as a percentage (from 1% to 100%) or by indicating a structure as a scaffold and then search for similar structures at different levels. Furthermore, this platform allows searching for structures with the same scaffold using different search algorithms. After several searches and attempts, the method with which it was possible to find more compounds with similar structures to compounds used to develop QSAR model 32 was by using the Ullmann algorithm.

By using this algorithm, natural compounds with four different scaffolds similar to the structural scaffold of the compounds present in G4 were searched. These scaffolds can be seen in **Figure 18**. After completing the search, a total of 1,628 compounds were collected into the library of test compounds.

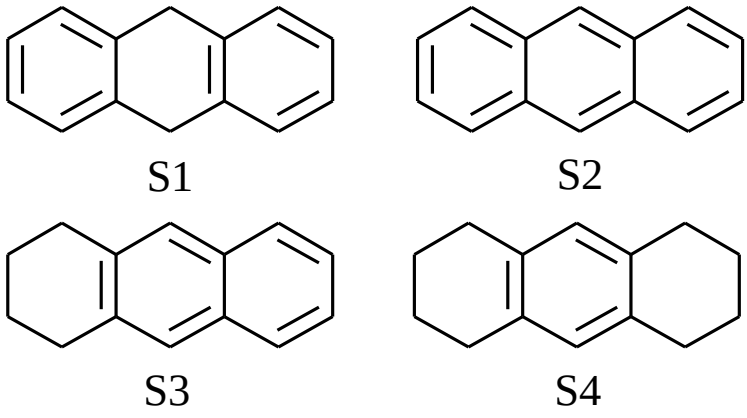


Figure 18 – Structures used as scaffolds for searching the COCONUT database.

All collected compounds can be viewed via the http://esa.ipb.pt/qsar/ website. On this website developed by the research team, a list of all compounds can be observed as well as the activity predicted by the QSAR model for each compound. Every compound is directly linked to the COCONUT platform providing all information available in this database relative to each compound.

It is possible to verify that the scaffold whose search resulted in more compounds for our test library was S1 containing 1,031 compounds, followed by S2, with 281 compounds, S3 with 206 compounds, and finally, S4 with 110 compounds.

Considering the structures of G4 compounds (**Figure 14-16**), used to build the QSAR model 32, this distribution seems logical, with S1 being the scaffold most similar to the compounds with the highest number present in G4. Since, as can be seen from **Table 3**, the models using G4 obtained, in general, good statistical data, it is possible to infer that compounds with an S1 scaffold are likely to have a strong relationship with the ability to inhibit *hs*TYR.

Through the analysis of the tested compounds, it is possible to verify that QSAR model 32 predicted that the 1 628 compounds have values for pIC_{50} between 2.77 and 6.27. For example, compound 608 was the one with the best predicted *hs*TYR inhibition ability, with a pIC_{50} value of 6.2683; corresponding to an IC_{50} concentration of 0.5391μM.

After predicting the *hs*TYR inhibiting activity of all compounds collected from the COCONUT database, the ZINC15 database was used to find and eventually purchase the most potent compounds. All compounds with a $pIC_{50\ PREDICTED}$ greater than 5.000 were considered for commercial acquisition, totaling 353 of the 1,628 compounds tested. Of all these compounds, only 18 different compounds were set for purchase and are in the process of being acquired. The purchased compounds can be seen in **Table 6**. As soon as the purchased compounds are delivered, they will be tested as potential *hs*TYR inhibitors and thus experimentally validate QSAR model 32.

Table 6 – Natural compounds purchased and to be tested apon arrival.

COMPOUND Nº	pIC_{50} PREDICTED	COCONUT ID	CAS Nº	ENTERPRISE ID	SUPLIER
648	5.7827	CNP0439104	3443-90-1	MolPort-003-909-021	AK Scientific, Inc.
293	5.1029	CNP0475182	3443-92-3	MolPort-006-116-200	AK Scientific, Inc.
344	5.2592	CNP0474656	67893-47-4	MolPort-006-117-940	AK Scientific, Inc.
367	5.1370	CNP0046760	No CAS	MolPort-002-139-351 MCULE-8967606569	ChemBridge Corporation
665	5.3727	CNP0052828	No CAS	MolPort-001-913-470 MCULE-8834371025	ChemDiv, Inc.
454	5.1085	CNP0035287	No CAS	MolPort-007-550-577 MCULE-2534118453	ChemDiv, Inc.
450	5.1626	CNP0054520	No CAS	MolPort-007-550-578 MCULE-3394290773	ChemDiv, Inc.
331	5.3251	CNP0336430	No CAS	MolPort-000-628-430 MCULE-2659259858	Vitas M Chemical Limited
338	5.5672	CNP0160897	No CAS	MolPort-000-628-431 MCULE-4610965051	Vitas M Chemical Limited
264	5.4767	CNP0160170	81-39-0	MolPort-000-628-457 MCULE-8543235543	Vitas M Chemical Limited
329	5.4312	CNP0370520	No CAS	MolPort-000-644-441	Vitas M Chemical Limited
257	5.4607	CNP0369027	No CAS	MolPort-000-645-394	Vitas M Chemical Limited
340	5.2265	CNP0328895	No CAS	MolPort-000-846-637 MCULE-7116644768	Vitas M Chemical Limited
377	5.0707	CNP0204913	No CAS	MolPort-001-014-913 MCULE-4737231165	Vitas M Chemical Limited
609	5.2573	CNP0362893	No CAS	MolPort-002-136-204	Vitas M Chemical Limited
332	5.3571	CNP0365332	No CAS	MolPort-002-363-875	Vitas M Chemical Limited
464	5.4682	CNP0468374	128-80-3	MolPort-002-685-168 MCULE-3441174162	Vitas M Chemical Limited
636	5.9438	CNP0412420	No CAS	MolPort-044-180-836	Vitas M Chemical Limited

4. Conclusion

An *in silico* library of 196 *hs*TYR inhibitors was prepared based on a literature search of experimental studies. From these compounds, it was possible to implement QSAR models of *hs*TYR activity. Compounds from the library were divided into 6 groups according to structural similarities. Molecular descriptors were calculated using several tools, testing the capabilities and usefulness of these descriptor calculation platforms. A total of 33 QSAR models were built using various combinations between the defined groups and the molecular descriptors calculated by different tools.

After analyzing all QSAR model's statistical data, QSAR model 32 was selected for further studies. This model used the G4 group composed of 41 compounds and the OCHEM descriptor calculation platform with a total of 28,933 molecular descriptors. These descriptors were subjected to statistical treatment to reduce the information, and the initial molecular descriptors were reduced to 3,126 and introduced into the PyQSAR tool. After analyzing the molecular descriptors provided, PyQSAR selected 4, using mathematical, statistical, and clustering methods, and a QSAR equation was obtained. The molecular descriptors selected were: C-026:(Dragon7), DISSM2C :(Mera), MaxdssC:(alvaDesc) and WHALES90_Rem:(alvaDesc). These molecular descriptors have a weak correlation with each other, demonstrating that each descriptor determines a different characteristic, thus reducing the probability of correlated values. In general, QSAR model 32 presented excellent statistical values for the determination coefficients (**R^2**=0.9128 and **R^2CV**=0.9147), for the mean error values (**RMSE**=0.1708 and **RMSE CV**=0.1878) and the mean score of the multiple linear regression method (**Q^2CV**=0.8922), confirming the good predictive power of the model.

A library of natural compounds with a structure similar to G4 compounds was prepared with the help of the search tools present in the COCONUT database of natural compounds, namely the method of searching for similar structures using the Ullmann algorithm. A total of 1,628 natural compounds were selected using 4 different structures as scaffolds for the compound research. The molecular descriptors of these compounds were calculated using OCHEM, and the equation obtained by the QSAR model was applied in order to predict the *hs*TYR inhibition capacity of each compound. The results are displayed on a website built by the research team and can be viewed by accessing the URL http://esa.ipb.pt/qsar/.

After predicting the inhibitory activity of the 1,628 in the natural compounds library, the ZINC15 database was used to verify which of these natural compounds were available for acquisition. Most of the 353 compounds searched were not available for purchase. A decision was made, and only 18 different compounds were set to buy. Upon arrival, these compounds will be tested against *hs*TYR to obtain their experimental capacity to inhibit this enzyme and thus validate QSAR model 32 as a predictive tool.

If one or more of the 18 compounds purchased present strong *hs*TYR inhibition activity, they would be considered for potential use in cosmeceutical applications related to an excess of melanin production, including skin-whitening and anti-pigmentation disorders applications.

5. References

[1] Simon, J. D., Peles, D., Wakamatsu, K., & Ito, S. (2009). Current challenges in understanding melanogenesis: bridging chemistry, biological control, morphology, and function. Pigment cell & melanoma research, 22(5), 563-579.

[2] Schiaffino, M. V. (2010). Signaling pathways in melanosome biogenesis and pathology. The international journal of biochemistry & cell biology, 42(7), 1094-1104.

[3] Yamaguchi, Y., Brenner, M., & Hearing, V. J. (2007). The regulation of skin pigmentation. Journal of biological chemistry, 282(38), 27557-27561.

[4] Sitaram, A., & Marks, M. S. (2012). Mechanisms of protein delivery to melanosomes in pigment cells. Physiology, 27(2), 85-99.

[5] Bonaventure, J., Domingues, M. J., & Larue, L. (2013). Cellular and molecular mechanisms controlling the migration of melanocytes and melanoma cells. Pigment cell & melanoma research, 26(3), 316-325.

[6] Winslow, T. (2008). Skin Anatomy. Terese Winslow LLC Medical Illustration.

[7] Passeron, T., Coelho, S. G., Miyamura, Y., Takahashi, K., & Hearing, V. J. (2007). Immunohistochemistry and in situ hybridization in the study of human skin melanocytes. Experimental dermatology, 16(3), 162-170.

[8] Hearing, V. J. (2011). Determination of melanin synthetic pathways. The Journal of investigative dermatology, 131(E1), E8.

[9] Lai, X., Wichers, H. J., Soler-Lopez, M., & Dijkstra, B. W. (2018). Structure and function of human tyrosinase and tyrosinase-related proteins. Chemistry–A European Journal, 24(1), 47-55.

[10] Roulier, B., Pérès, B., & Haudecoeur, R. (2020). Advances in the design of genuine human tyrosinase inhibitors for targeting melanogenesis and related pigmentations. Journal of Medicinal Chemistry, 63(22), 13428-13443.

[11] Fogal, S., Carotti, M., Giaretta, L., Lanciai, F., Nogara, L., Bubacco, L., & Bergantino, E. (2015). Human tyrosinase produced in insect cells: a landmark for the screening of new drugs addressing its activity. Molecular biotechnology, 57(1), 45-57.

[12] Dolinska, M. B., Wingfield, P. T., Young, K. L., & Sergeev, Y. V. (2019). The TYRP1-mediated protection of human tyrosinase activity does not involve stable interactions of tyrosinase domains. Pigment cell & melanoma research, 32(6), 753-765.

[13] National Center for Biotechnology Information (2022). PubChem Pathway Summary for Pathway SMP0000006, Tyrosine Metabolism, Source: PathBank. Retrieved January 22, 2022 from https://pubchem.ncbi.nlm.nih.gov/pathway/PathBank:SMP0000006.

[14] Chang, T. S. (2012). Natural melanogenesis inhibitors acting through the down-regulation of tyrosinase activity. Materials, 5(9), 1661-1685.

[15] Costin, G. E., & Hearing, V. J. (2007). Human skin pigmentation: melanocytes modulate skin color in response to stress. The FASEB journal, 21(4), 976-994.

[16] Buitrago, E., Hardre, R., Haudecoeur, R., Jamet, H., Belle, C., Boumendjel, A., Bubacco, L., & Reglier, M. (2016). Are human tyrosinase and related proteins suitable targets for melanoma therapy?. Current topics in medicinal chemistry, 16(27), 3033-3047.

[17] Naidoo, L., Khoza, N., & Dlova, N. C. (2016). A fairer face, a fairer tomorrow? A review of skin lighteners. Cosmetics, 3(3), 33.

[18] Burger, P., Landreau, A., Azoulay, S., Michel, T., & Fernandez, X. (2016). Skin whitening cosmetics: Feedback and challenges in the development of natural skin lighteners. Cosmetics, 3(4), 36.

[19] Sagoe, D., Pallesen, S., Dlova, N. C., Lartey, M., Ezzedine, K., & Dadzie, O. (2019). The global prevalence and correlates of skin bleaching: a meta-analysis and meta-regression analysis. International journal of dermatology, 58(1), 24-44.

[20] The Global Coalition for Melanoma Patient Advocacy (2020). Melanoma Skin Cancer Report. https://melanomapatients.org.au (accessed Jan 11, 2022).

[21] McDermott, D., Lebbé, C., Hodi, F. S., Maio, M., Weber, J. S., Wolchok, J. D., Thompson, J., & Balch, C. M. (2014). Durable benefit and the potential for long-term survival with immunotherapy in advanced melanoma. Cancer treatment reviews, 40(9), 1056-1064.

[22] Espenel, S., Vallard, A., Rancoule, C., Garcia, M. A., Guy, J. B., Chargari, C., Deutsch, E., & Magné, N. (2017). Melanoma: last call for radiotherapy. Critical reviews in oncology/hematology, 110, 13-19.

[23] Haining, R. L., & Achat-Mendes, C. (2017). Neuromelanin, one of the most overlooked molecules in modern medicine, is not a spectator. Neural regeneration research, 12(3), 372.

[24] Zucca, F. A., Segura-Aguilar, J., Ferrari, E., Muñoz, P., Paris, I., Sulzer, D., Sarna, T., Casella, L., & Zecca, L. (2017). Interactions of iron, dopamine and neuromelanin pathways in brain aging and Parkinson's disease. Progress in neurobiology, 155, 96-119.

[25] Bose, A., Petsko, G. A., & Eliezer, D. (2018). Parkinson's disease and melanoma: co-occurrence and mechanisms. Journal of Parkinson's disease, 8(3), 385-398.

[26] Saruno, R., Kato, F., & Ikeno, T. (1979). Kojic acid, a tyrosinase inhibitor from Aspergillus albus. Agricultural and Biological Chemistry, 43(6), 1337-1338.

[27] Mann, T., Gerwat, W., Batzer, J., Eggers, K., Scherner, C., Wenck, H., Stäb, F., Hearing, V. J., Röhm, K. H., & Kolbe, L. (2018). Inhibition of human tyrosinase requires molecular motifs distinctively different from mushroom tyrosinase. Journal of Investigative Dermatology, 138(7), 1601-1608.

[28] Ranganath, L. R., Norman, B. P., & Gallagher, J. A. (2019). Ochronotic pigmentation is caused by homogentisic acid and is the key event in alkaptonuria leading to the destructive consequences of the disease—a review. Journal of inherited metabolic disease, 42(5), 776-792.

[29] Mann, T., Scherner, C., Röhm, K. H., & Kolbe, L. (2018). Structure-activity relationships of thiazolyl resorcinols, potent and selective inhibitors of human tyrosinase. International journal of molecular sciences, 19(3), 690.

[30] Gini, G. (2018). QSAR: What Else?. In Computational Toxicology (pp. 79-105). Humana Press, New York, NY.

[31] Goya Jorge E., Rayar A. M., Barigye S. J., Jorge Rodríguez M. E., & Sylla-Iyarreta Veitía M. (2016). Development of an *in silico* Model of DPPH• Free Radical Scavenging Capacity: Prediction of Antioxidant Activity of Coumarin Type Compounds. International Journal of Molecular Sciences, 17(6), 881.

[32] Hansch C., Maloney P. P., Fujita T., & Muir R. M. (1962). Correlation of biological Activity of phenoxyacetic acids with Hammett substituent constants and partition coefficients. Nature, 194(4824), 178-180.

[33] Roy K., Kar S., & Das R. N. (2015). "Chapter 1.2: What is QSAR? Definitions and Formulism". A primer on QSAR/QSPR modeling: Fundamental Concepts. New York: Springer-Verlag Inc., 2–6.

[34] Ghasemi F., Mehridehnavi A., Perez-Garrido A., & Perez-Sanchez H. (2018). Neural network and deep-learning algorithms used in QSAR studies: merits and drawbacks. Drug Discov. Today, 23(10), 1784-1790.

[35] Xue L., & Bajorath J. (2000). Molecular descriptors in chemoinformatics, computational combinatorial chemistry, and virtual screening. Combinatorial chemistry & high throughput screening, 3(5), 363-372.

[36] Hopfinger A. J., Wang S., Tokarski J. S., Jin B., Albuquerque M., Madhav P. J., & Duraiswami C. (1997). Construction of 3D-QSAR models using the 4D-QSAR analysis formalism. Journal of the American Chemical Society, 119(43), 10509-10524.

[37] Vedani A., & Dobler M. (2002). 5D-QSAR: the key for simulating induced fit?. Journal of medicinal chemistry, 45(11), 2139-2149.

[38] Vedani A., Dobler M., & Lill M. A. (2005). Combining protein modeling and 6D-QSAR. Simulating the binding of structurally diverse ligands to the estrogen receptor. Journal of medicinal chemistry, 48(11), 3700-3703.

[39] Todeschini R., Consonni V., & Mannhold R. (2000). Methods and principles in medicinal chemistry. Kubinyi H, Timmerman H (Series eds) Handbook of molecular descriptors. Wiley-VCH, Weinheim.

[40] Oliveira, N., Abreu, R., & Adega, F. (2021). Tirosinase humana como proteína-alvo para doenças de pele: estudos de modelação molecular. Relatório Final de Estágio de Licenciatura em Biologia na Universidade de Trás-os-Montes e Alto Douro (UTAD).

[41] Nguyen, M. T., Le, T. H., Nguyen, H. X., Dang, P. H., Do, T. N., Abe, M., Takagi, R., & Nguyen, N. T. (2017). Artocarmins G–M, prenylated 4-chromenones from the stems of Artocarpus rigida and their tyrosinase inhibitory activities. Journal of natural products, 80(12), 3172-3178.

[42] Rosa, G. P., Palmeira, A., Almeida, I. F., Kane-Pagès, A., Barreto, M. C., Sousa, E., & Pinto, M. M. M. (2021). Xanthones for melanogenesis inhibition: Molecular docking and QSAR studies to understand their anti-tyrosinase activity. Bioorganic & Medicinal Chemistry, 29, 115873.

[43] Wu, Y., Wu, Z. R., Chen, P., Deng, W. R., Wang, Y. Q., & Li, H. Y. (2015). Effect of the tyrosinase inhibitor (S)-N-trans-feruloyloctopamine from garlic skin on tyrosinase gene expression and melanine accumulation in melanoma cells. Bioorganic & Medicinal Chemistry Letters, 25(7), 1476-1478.

[44] Casañola-Martín, G. M., Khan, M. T. H., Marrero-Ponce, Y., Ather, A., Sultankhodzhaev, M. N., & Torrens, F. (2006). New tyrosinase inhibitors selected by atomic linear indices-based classification models. Bioorganic & medicinal chemistry letters, 16(2), 324-330.

[45] Young, S. S., Yuan, F., & Zhu, M. (2012). Chemical descriptors are more important than learning algorithms for modelling. Molecular informatics, 31(10), 707-710.

[46] Hongmao, S. (2016). Chapter 5-Quantitative Structure–Activity Relationships: Promise, Validations, and Pitfalls. A Practical Guide to Rational Drug Design. Woodhead Publishing, 163-192.

[47] Sushko, I., Novotarskyi, S., Körner, R., Pandey, A. K., Rupp, M., Teetz, W., Brandmaier, S., Abdelaziz, A., Prokopenko, V. V., Tanchuk, V. Y., Todeschini, R., Varnek, A., Marcou, G., Ertl, P., Potemkin, V., Grishina, M., Gasteiger, J., Schwab, C., Baskin, I. I., Palyulin, V. A., Radchenko, E. V., Welsh, W. J., Kholodovych, V., Chekmarev, D., Cherkasov, A., Aires-de-Sousa, J., Zhang, Q. Y., Bender, A., Nigsch, F., Patiny, L., Williams, A., Tkachenko, V., Tetko, I. V. (2011). Online chemical modeling environment (OCHEM): web platform for data storage, model development and publishing of chemical information. Journal of computer-aided molecular design, 25(6), 533-554.

[48] A Gentle Introduction to k-fold Cross-Validation (2018). Machine Learning Mastery. https://machinelearningmastery.com/k-fold-cross-validation/ (accessed Sep 10, 2022).

[49] Sterling, T., & Irwin, J. J. (2015). ZINC 15–ligand discovery for everyone. Journal of chemical information and modeling, 55(11), 2324-2337.

[50] Sorokina, M., Merseburger, P., Rajan, K., Yirik, M. A., & Steinbeck, C. (2021). COCONUT online: collection of open natural products database. Journal of Cheminformatics, 13(1), 1-13.

[51] COCONUT (COlleCtion of Open Natural prodUcTs). https://coconut.naturalproducts.net/documentation (accessed Sep 20, 2022).

[52] Moonrungsee, N., Shimamura, T., Kashiwagi, T., Jakmunee, J., Higuchi, K., & Ukeda, H. (2012). Sequential injection spectrophotometric system for evaluation of mushroom tyrosinase-inhibitory activity. Talanta, 101, 233-239.

[53] Gao, H. (2018). Predicting tyrosinase inhibition by 3D QSAR pharmacophore models and designing potential tyrosinase inhibitors from Traditional Chinese medicine database. Phytomedicine, 38, 145-157.

[54] Li, K., Ji, S., Song, W., Kuang, Y., Lin, Y., Tang, S., Cui, Z., Qiao, X., Yu, S,. & Ye, M. (2017). Glycybridins A–K, bioactive phenolic compounds from Glycyrrhiza glabra. Journal of Natural Products, 80(2), 334-346.

[55] Todeschini, R., & Consonni, V. (2009). Molecular Descriptors for Chemoinformatics, 2 Volume Set: Volume I: Alphabetical Listing/Volume II: Appendices, References (Vol. 41). Wiley-VCH.

[56] OCHEM (Online chemical database). https://docs.ochem.eu/display/MAN/MERA+descriptors.html (accessed Oct 13, 2022).

[57] DEDuCT (Database of endocrine disrupting chemicals and their toxicity profiles). https://cb.imsc.res.in/deduct/descriptors/eJaFhpBpbGtp (accessed Oct 13, 2022).

[58] Grisoni, F., Merk, D., Consonni, V., Hiss, J. A., Tagliabue, S. G., Todeschini, R., & Schneider, G. (2018). Scaffold hopping from natural products to synthetic mimetics by holistic molecular similarity. Communications Chemistry, 1(1), 1-9.

[59] Consonni, V., Ballabio, D., & Todeschini, R. (2009). Comments on the definition of the Q 2 parameter for QSAR validation. Journal of chemical information and modeling, 49(7), 1669-1678.

[60] Panzella, L., & Napolitano, A. (2019). Natural and bioinspired phenolic compounds as tyrosinase inhibitors for the treatment of skin hyperpigmentation: Recent advances. Cosmetics, 6(4), 57.

[61] Haudecoeur, R., Carotti, M., Gouron, A., Maresca, M., Buitrago, E., Hardré, R., Bergantino, E., Jamet, H., Belle, C., Réglier, M., Bubacco, L., & Boumendjel, A. (2017). 2-Hydroxypyridine-n-oxide-embedded aurones as potent human tyrosinase inhibitors. ACS medicinal chemistry letters, 8(1), 55-60.

[62] Okombi, S., Rival, D., Bonnet, S., Mariotte, A. M., Perrier, E., & Boumendjel, A. (2006). Discovery of benzylidenebenzofuran-3 (2 H)-one (aurones) as inhibitors of tyrosinase derived from human melanocytes. Journal of medicinal chemistry, 49(1), 329-333.

[63] Ji, S., Li, Z., Song, W., Wang, Y., Liang, W., Li, K., Tang, S., Wang, Q., Qiao, X., Zhou, D., Yu, S., & Ye, M. (2016). Bioactive constituents of Glycyrrhiza uralensis (licorice): discovery of the effective components of a traditional herbal medicine. Journal of Natural Products, 79(2), 281-292.

[64] Pillaiyar, T., Namasivayam, V., Manickam, M., & Jung, S. H. (2018). Inhibitors of melanogenesis: an updated review. Journal of medicinal chemistry, 61(17), 7395-7418.

[65] Takara, K., Iwasaki, H., Ujihara, K., & Wada, K. (2008). Human tyrosinase inhibitor in rum distillate wastewater. Journal of Oleo Science, 57(3), 191-196.

[66] Likhitwitayawuid, K., Sornsute, A., Sritularak, B., & Ploypradith, P. (2006). Chemical transformations of oxyresveratrol (trans-2, 4, 3′, 5′-tetrahydroxystilbene) into a potent tyrosinase inhibitor and a strong cytotoxic agent. Bioorganic & medicinal chemistry letters, 16(21), 5650-5653.

[67] Iwadate, T., & Nihei, K. I. (2015). Rhododendrol glycosides as stereospecific tyrosinase inhibitors. Bioorganic & Medicinal Chemistry, 23(20), 6650-6658.

[68] Lee, C. W., Son, E. M., Kim, H. S., Xu, P., Batmunkh, T., Lee, B. J., & Koo, K. A. (2007). Synthetic tyrosyl gallate derivatives as potent melanin formation inhibitors. Bioorganic & medicinal chemistry letters, 17(19), 5462-5464.

[69] Catalano, M., Bassi, G., Rotondi, G., Khettabi, L., Dichiara, M., Murer, P., Scheuermann, J., Soler-Lopez, M., & Neri, D. (2021). Discovery, affinity maturation and multimerization of small molecule ligands against human tyrosinase and tyrosinase-related protein 1. RSC medicinal chemistry, 12(3), 363-369.

[70] Husain, A., Khan, S. A., Iram, F., Iqbal, M. A., & Asif, M. (2019). Insights into the chemistry and therapeutic potential of furanones: A versatile pharmacophore. European Journal of Medicinal Chemistry, 171, 66-92.

[71] Yoshimori, A., Oyama, T., Takahashi, S., Abe, H., Kamiya, T., Abe, T., & Tanuma, S. I. (2014). Structure–activity relationships of the thujaplicins for inhibition of human tyrosinase. Bioorganic & medicinal chemistry, 22(21), 6193-6200.

[72] Sugimoto, K., Nishimura, T., Nomura, K., Sugimoto, K., & Kuriki, T. (2003). Syntheses of arbutin-α-glycosides and a comparison of their inhibitory effects with those of α-arbutin and arbutin on human tyrosinase. Chemical and pharmaceutical bulletin, 51(7), 798-801.

[73] Sugimoto, K., Nomura, K., Nishimura, T., Kiso, T., Sugimoto, K., & Kuriki, T. (2005). Syntheses of α-arbutin-α-glycosides and their inhibitory effects on human tyrosinase. Journal of bioscience and bioengineering, 99(3), 272-276.

[74] Sugimoto, S., Yamano, Y., Desoukey, S. Y., Katakawa, K., Wanas, A. S., Otsuka, H., & Matsunami, K. (2019). Isolation of Sesquiterpene–Amino Acid Conjugates, Onopornoids A–D, and a Flavonoid Glucoside from Onopordum alexandrinum. Journal of natural products, 82(6), 1471-1477.

[75] Ishioka, W., Oonuki, S., Iwadate, T., & Nihei, K. I. (2019). Resorcinol alkyl glucosides as potent tyrosinase inhibitors. Bioorganic & Medicinal Chemistry Letters, 29(2), 313-316.

[76] Su, C. R., Kuo, P. C., Wang, M. L., Liou, M. J., Damu, A. G., & Wu, T. S. (2003). Acetophenone Derivatives from Acronychia p edunculata. Journal of natural products, 66(7), 990-993.

[77] Cho, S., Kim, S. H., & Shin, D. (2019). Recent applications of hydantoin and thiohydantoin in medicinal chemistry. European journal of medicinal chemistry, 164, 517-545.

6. Supporting Information

Table S1 – Transformation of IC_{50} values into pIC_{50} of compounds present in the compound library.

COMPOUND NUMBER		IC_{50} (μM)	pIC_{50}	COMPOUND NUMBER		IC_{50} (μM)	pIC_{50}
GROUP 1	**1**[53]	2,720	5,565	GROUP 2	**54**[64]	19,200	4,717
	2[53]	0,110	6,959		**55**[64]	0,070	7,155
	3[53]	1,960	5,708	GROUP 3	**56**[53]	6,130	5,213
	4[60]	20,000	4,699		**57**[53]	20,500	4,688
	5[41]	0,023	7,638		**58**[53]	58,060	4,236
	6[61]	30,000	4,523		**59**[53]	59,700	4,224
	7[54]	7,500	5,125		**60**[53]	20,100	4,697
	8[61]	85,300	4,069		**61**[53]	27,100	4,567
	9[54]	0,090	7,046		**62**[53]	54,400	4,264
	10[41]	45,300	4,344		**63**[29]	3,200	5,495
	11[62]	31,700	4,499		**64**[29]	19,000	4,721
	12[62]	38,400	4,416		**65**[29]	750,000	3,125
	13[10]	16,000	4,796		**66**[29]	56,000	4,252
	14[63]	0,177	6,752		**67**[29]	3,500	5,456
	15[63]	0,154	6,812		**68**[29]	15,000	4,824
	16[63]	1,000	6,000		**69**[29]	33,000	4,481
	17[64]	0,980	6,009		**70**[29]	5,600	5,252
	18[61]	119,000	3,924		**71**[29]	6,200	5,208
	19[10]	60,000	4,222		**72**[29]	25,000	4,602
	20[64]	290,000	3,538		**73**[29]	60,000	4,222
	21[64]	8,000	5,097		**74**[29]	16,000	4,796
	22[62]	38,000	4,420		**75**[29]	10,000	5,000
	23[54]	5,100	5,292		**76**[29]	5,700	5,244
GROUP 2	**24**[53]	0,029	7,538		**77**[29]	81,000	4,092
	25[53]	250,000	3,602		**78**[29]	9,800	5,009
	26[53]	250,000	3,602		**79**[29]	6,900	5,161
	27[53]	79,050	4,102		**80**[29]	140,000	3,854
	28[53]	250,000	3,602		**81**[29]	160,000	3,796
	29[53]	250,000	3,602		**82**[29]	3,500	5,456
	30[65]	18500,00	1,733		**83**[29]	1400,000	2,854
	31[10]	1,700	5,770		**84**[29]	40,000	4,398
	32[64]	4,770	5,321		**85**[10]	3,200	5,495
	33[64]	12,600	4,900		**86**[10]	1,600	5,796
	34[10]	40,000	4,398		**87**[10]	4,600	5,337
	35[64]	1,950	5,710		**88**[10]	1,400	5,854
	36[10]	5,000	5,301		**89**[10]	2,500	5,602
	37[66]	12,700	4,896		**90**[10]	1,100	5,959
	38[10]	2,000	5,699		**91**[70]	5,600	5,252
	39[64]	7,890	5,103		**92**[10]	2,600	5,585
	40[66]	1,600	5,796		**93**[10]	3,200	5,495
	41[67]	4,560	5,341		**94**[10]	51,000	4,292
	42[68]	15,210	4,818	GROUP 4	**95**[53]	28,050	4,552
	43[64]	0,080	7,097		**96**[53]	58,060	4,236
	44[10]	9,100	5,041		**97**[53]	90,990	4,041
	45[68]	14,500	4,839		**98**[53]	108,620	3,964
	46[10]	85,000	4,071		**99**[53]	200,000	3,699
	47[10]	141,000	3,851		**100**[53]	200,000	3,699
	48[10]	2,500	5,602		**101**[53]	24,520	4,610
	49[68]	4,930	5,307		**102**[53]	32,810	4,484
	50[64]	0,170	6,770		**103**[53]	83,170	4,080
	51[10]	32,000	4,495		**104**[53]	200,000	3,699
	52[10]	10,000	5,000		**105**[54]	2,800	5,553
	53[10]	50,000	4,301		**106**[42]	8,930	5,049

Table S1 (continuation) – Transformation of IC_{50} values into pIC_{50} of compounds present in the compound library.

COMPOUND NUMBER		IC_{50} (μM)	pIC_{50}	COMPOUND NUMBER		IC_{50} (μM)	pIC_{50}
GROUP 4	**107**[42]	3,160	5,500	GROUP 5	**152**[10]	1,000	6,000
	108[42]	4,520	5,345		**153**[67]	1,680	5,775
	109[42]	89,370	4,049		**154**[64]	21,000	4,678
	110[42]	8,290	5,081		**155**[67]	2,170	5,664
	111[42]	5,200	5,284		**156**[71]	1,150	5,939
	112[42]	7,800	5,108		**157**[10]	94,000	4,027
	113[42]	8,470	5,072		**158**[10]	22,000	4,658
	114[42]	4,020	5,396		**159**[41]	9,350	5,029
	115[42]	4,420	5,355		**160**[10]	10,000	5,000
	116[42]	33,700	4,472		**161**[10]	1,000	6,000
	117[42]	22,400	4,650		**162**[64]	8,980	5,047
	118[42]	29,130	4,536		**163**[10]	76,000	4,119
	119[42]	5,120	5,291	GROUP 6	**164**[53]	1,670	5,777
	120[42]	5,600	5,252		**165**[53]	8,610	5,065
	121[42]	3,010	5,521		**166**[53]	25,810	4,588
	122[42]	5,120	5,291		**167**[53]	63,500	4,197
	123[42]	5,570	5,254		**168**[53]	16,170	4,791
	124[42]	3,810	5,419		**169**[53]	16,900	4,772
	125[42]	1,900	5,721		**170**[72]	5700,000	2,244
	126[42]	2,000	5,699		**171**[72]	6100,000	2,215
	127[42]	5,700	5,244		**172**[73]	2,100	5,678
	128[42]	5,330	5,273		**173**[73]	4900,000	2,310
	129[42]	9,260	5,033		**174**[73]	13900,00	1,857
	130[42]	8,840	5,054		**175**[67]	1,980	5,703
	131[42]	5,140	5,289		**176**[67]	1,510	5,821
	132[42]	8,830	5,054		**177**[44]	102,390	3,990
	133[42]	3,280	5,484		**178**[44]	54,640	4,262
	134[42]	5,140	5,289		**179**[74]	50,000	4,301
	135[42]	8,460	5,073		**180**[74]	50,000	4,301
GROUP 5	**136**[53]	200,000	3,699		**181**[44]	48,920	4,311
	137[53]	300,000	3,523		**182**[68]	30,260	4,519
	138[53]	300,000	3,523		**183**[67]	1,720	5,764
	139[29]	650,000	3,187		**184**[64]	4,620	5,335
	140[29]	220,000	3,658		**185**[75]	417,000	3,380
	141[10]	1,700	5,770		**186**[67]	4,130	5,384
	142[10]	3000,000	2,523		**187**[44]	85,010	4,071
	143[10]	2,000	5,699		**188**[67]	3,830	5,417
	144[43]	54,200	4,266		**189**[67]	2,300	5,638
	145[43]	9,100	5,041		**190**[76]	333,000	3,478
	146[10]	30,000	4,523		**191**[64]	8,970	5,047
	147[43]	32,500	4,488		**192**[77]	7,360	5,133
	148[64]	30,000	4,523		**193**[77]	1,070	5,971
	149[64]	4400,000	2,357		**194**[44]	13,950	4,855
	150[70]	50,000	4,301		**195**[44]	92,250	4,035
	151[61]	150,000	3,824		**196**[67]	4,720	5,326