UNIVERSIDAD DE INVESTIGACIÓN DE TECNOLOGÍA EXPERIMENTAL YACHAY

Escuela de Ciencias Biológicas e Ingeniería

TÍTULO: Fishing Potent Tyrosinase Inhibitors with New Molecular Docking Protocol from Natural Products Ocean

Trabajo de integración curricular presentado como requisito para la obtención del título de Bióloga

Autora:

Camila Lissett Velastegui Gamboa

Tutor:

Ph.D. Nelson Santiago

Co-tutor:

Ph.D. Yovani Marrero – Ponce

Urcuquí, Marzo 2021.

Urcuquí, 7 de junio de 2021

SECRETARÍA GENERAL
(Vicerrectorado Académico/Cancillería)
ESCUELA DE CIENCIAS BIOLÓGICAS E INGENIERÍA
CARRERA DE BIOLOGÍA
ACTA DE DEFENSA No. UITEY-BIO-2021-00010-AD

A los 7 días del mes de junio de 2021, a las 09:30 horas, de manera virtual mediante videoconferencia, y ante el Tribunal Calificador, integrado por los docentes:

Presidente Tribunal de Defensa	Dr. SINCHE CHELE, FEDERICO LEONARDO , Ph.D.
Miembro No Tutor	Dr. ALEXIS FRANK , Ph.D.
Tutor	Dr. SANTIAGO VISPO, NELSON FRANCISCO , Ph.D.

El(la) señor(ita) estudiante **VELASTEGUI GAMBOA, CAMILA LISSETT**, con cédula de identidad No. **1805000773**, de la **ESCUELA DE CIENCIAS BIOLÓGICAS E INGENIERÍA**, de la Carrera de **BIOLOGÍA**, aprobada por el Consejo de Educación Superior (CES), mediante Resolución **RPC-SO-37-No.438-2014**, realiza a través de videoconferencia, la sustentación de su trabajo de titulación denominado: **Fishing Potent Tyrosinase Inhibitors with New Molecular Docking Protocol from Natural Products Ocean** , previa a la obtención del título de **BIÓLOGO/A**.

El citado trabajo de titulación, fue debidamente aprobado por el(los) docente(s):

Tutor	Dr. SANTIAGO VISPO, NELSON FRANCISCO , Ph.D.

Y recibió las observaciones de los otros miembros del Tribunal Calificador, las mismas que han sido incorporadas por el(la) estudiante.

Previamente cumplidos los requisitos legales y reglamentarios, el trabajo de titulación fue sustentado por el(la) estudiante y examinado por los miembros del Tribunal Calificador. Escuchada la sustentación del trabajo de titulación a través de videoconferencia, que integró la exposición de el(la) estudiante sobre el contenido de la misma y las preguntas formuladas por los miembros del Tribunal, se califica la sustentación del trabajo de titulación con las siguientes calificaciones:

Tipo	Docente	Calificación
Miembro Tribunal De Defensa	Dr. ALEXIS FRANK , Ph.D.	10,0
Presidente Tribunal De Defensa	Dr. SINCHE CHELE, FEDERICO LEONARDO , Ph.D.	10,0
Tutor	Dr. SANTIAGO VISPO, NELSON FRANCISCO , Ph.D.	10,0

Lo que da un promedio de: **10 (Diez punto Cero)**, sobre 10 (diez), equivalente a: **APROBADO**

Para constancia de lo actuado, firman los miembros del Tribunal Calificador, el/la estudiante y el/la secretario ad-hoc.

Certifico que *en cumplimiento del Decreto Ejecutivo 1017 de 16 de marzo de 2020, la defensa de trabajo de titulación (o examen de grado modalidad teórico práctica) se realizó vía virtual, por lo que las firmas de los miembros del Tribunal de Defensa de Grado, constan en forma digital.*

VELASTEGUI GAMBOA, CAMILA LISSETT
Estudiante

FEDERICO LEONARDO SINCHE CHELE

Dr. SINCHE CHELE, FEDERICO LEONARDO , Ph.D.
Presidente Tribunal de Defensa

NELSON FRANCISCO SANTIAGO VISPO

Dr. SANTIAGO VISPO, NELSON FRANCISCO , Ph.D.
Tutor

Firmado electrónicamente por:
FRANK ALEXIS .

Dr. ALEXIS FRANK , Ph.D.
Miembro No Tutor

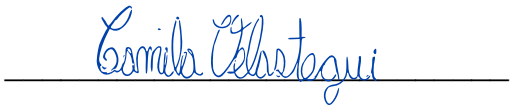
KARLA ESTEFANIA ALARCON FELIX
Firmado digitalmente por KARLA ESTEFANIA ALARCON FELIX
Fecha: 2021-08-07 11:52:57 -0500

ALARCON FELIX, KARLA ESTEFANIA
Secretario Ad-hoc

AUTORÍA

Yo, **Camila Lissett Velastegui Gamboa**, con cédula de identidad 180500077-3, declaro que las ideas, juicios, valoraciones, interpretaciones, consultas bibliográficas, definiciones y conceptualizaciones expuestas en el presente trabajo; así cómo, los procedimientos y herramientas utilizadas en la investigación, son de absoluta responsabilidad de el/la autora (a) del trabajo de integración curricular. Así mismo, me acojo a los reglamentos internos de la Universidad de Investigación de Tecnología Experimental Yachay.

Urcuquí, Marzo, 2021.



Camila Lisset Velastegui Gamboa

CI: 1805000773

AUTORIZACIÓN DE PUBLICACIÓN

Yo, **Camila Lissett Velastegui Gamboa**, con cédula de identidad 180500077-3, cedo a la Universidad de Tecnología Experimental Yachay, los derechos de publicación de la presente obra, sin que deba haber un reconocimiento económico por este concepto. Declaro además que el texto del presente trabajo de titulación no podrá ser cedido a ninguna empresa editorial para su publicación u otros fines, sin contar previamente con la autorización escrita de la Universidad.

Asimismo, autorizo a la Universidad que realice la digitalización y publicación de este trabajo de integración curricular en el repositorio virtual, de conformidad a lo dispuesto en el Art. 144 de la Ley Orgánica de Educación Superior

Urcuquí, Marzo, 2021.

Camila Lisset Velastegui Gamboa

CI: 1805000773

Dedicatoria

Para mis mis padres, Beatriz y Leonardo, que siempre me han acompañado, apoyado y amado incondicionalmente en todo. Y sin los cuales no lo habría logrado.

Para mis hermanos Leonardo, Fernanda y Carlos, que han cuidado de mi y me han alegrado los días.

-

Camila Lisset Velastegui Gamboa

Acknowledgments

-

I would like to thank my tutor Nelson Santiago, who since I met him guided me and advised me to continue during this biology journey. As well I want to thank my co-tutor Yovani Marrero Ponce who has taught me, guided me, and worked alongside me for the development of this research and without him, this work would not be possible. Special thanks to the group of programmers who implemented the software, and the personal of the Mathematical department that allowed me access to laboratory computers.

I want to extend a special thank you to all my professors for transmitting to me that love and dedication for science and learning that has always inspired me. I want to acknowledge as well my bio-friends Mary and Leandro that made my days better during the career. And finally thanks to my friends from Yachay Tech who become my family especially Jenny, Joha, Jonathan, and Daya that always supported me, taught me, and motivated me since the leveling course.

Camila Lisset Velastegui Gamboa

Resumen

La melanina tiene un rol biológico importante, aunque su sobreproducción y acumulación están asociadas a algunas enfermedades que van desde problemas cutáneos como el melasma hasta la enfermedad de Parkinson. La sobreproducción de melanina se debe a un aumento del número de melanocitos o de la actividad de las enzimas melanogénicas. Los melanocitos están regulados principalmente por la enzima tirosinasa que cataliza los dos pasos más importantes para la producción de pigmentos. Esto hace que la oxidorreductasa sea el principal objetivo para inhibir la producción de melanina. Actualmente, existen algunos inhibidores de la tirosinasa (TI) que se utilizan para el tratamiento de diversos trastornos de la piel. Sin embargo, los TI actuales estan asociados a algunos problemas por su alta citotoxicidad, mutagénesis, efectos no duraderos, entre otros. Por lo tanto, es clara la necesidad de Tis nuevos, más eficaces, de amplio espectro, más segura y más duradera efecto. Los métodos "*in vitro*" existentes para encontrar nuevos inhibidores son costosos y de eficacia limitada. El presente estudio se enfocará en desarrollar un estudio teórico con un protocolo de acoplamiento molecular que utiliza por primera vez un panel diverso de tirosinasas y permite el descubrimiento de nuevos TI utilizando un sistema experto. Aquí, desarrollamos modelos de clasificación con técnicas de aprendizaje automático basados en variables de panel de acoplamiento molecular de tirosinasa. Desarrollamos e integramos todo este flujo de trabajo en el software gratuito disponible SiLis-PREENZA para la predicción de TI. Además, hicimos una actualización de los inhibidores de tirosinasa conocidos de 701 a 2514 reportados hasta ahora. Este informe será útil para mejorar el proceso de búsqueda de inhibidores de tirosinasa para buscar nuevos agentes despigmentantes para el tratamiento de trastornos de hiperpigmentación. Los compuestos identificados como inhibidores podrían usarse para limitar los experimentos "in vitro" primero con experimentos fenotípicos hasta llegar a las células del melanoma.

Palabras Clave: Inhibidor de tirosinas, Acoplamiento Molecular, Autodock-Vina, Aprendizaje automatico, Modelos predictivos, WEKA, SiLis-PREENZA Software, Cribado virtual, Productos Naturales

Abstract

Melanin has an important biological role, although their overproduction and accumulation are associated with some diseases going from skin problems as melasma to Parkinson's disease. The overproduction of melanin is due to an increase in the number of melanocytes or melanogenic enzyme activity. The melanocytes are regulated mainly by the tyrosinase enzyme, which catalyzes the two most important steps to produce pigments. This makes the oxidoreductase the principal target to inhibit melanin production. Currently, there are some tyrosinase inhibitors (TIs) used for the treatment of various skin disorders.

Nevertheless, the current TIs had associated some safety problems such as their high cytotoxicity, mutagenesis, non-lasting effects, among others. Therefore, the need for new, more effective, broad-spectrum, safer, and longer-lasting protection of TIs is clear. The existing *"in vitro*" methods to find new inhibitors are expensive and with limited efficacy. The present study will focus on developing a theoretical study with a molecular docking protocol that uses a diverse panel of tyrosinases for the first time and allows the discovery of new TIs using an expert system. Here, we developed classification models with machine learning techniques based on the tyrosinase panel's molecular docking. We developed and integrated all this workflow into free available software SiLis PREENZA for the prediction of TIs.

Furthermore, we made an update of the known tyrosinase inhibitors of 701 to 2514 reported so far. This report will help enhance the process of finding tyrosinase inhibitors to seek novel depigmenting agents for the treatment of hyperpigmentation disorders. The compounds identified as inhibitors could be used to narrow "*in vitro"* experimentations first with phenotypic experiments until they reach melanoma cells.

Key Words: Tyrosinase Inhibitor, Molecular Docking, Autodock-Vina, Machine Learning, Predictive Model, WEKA, SiLis-PREENZA Software, Virtual Screening, Natural Product

Contents

List of Figures

List of Tables

PUBLICATION MANUSCRIPT

TITLE:

Fishing Potent Tyrosinase Inhibitors with New Molecular Docking Protocol from Natural Products Ocean

JOURNAL:

Journal of Chemical Information and Modeling -

AUTHORS:

Camila Lissett Velastegui Gamboa

Yovani Marrero-Ponce

Nelson Santiago Vispo

ORCID:

0000-0002-0340-6247

0000-0003-2721-1142

0000-0002-4081-3984

e-mail:

camila.velastegui@yachaytech.edu.ec

ymarrero@usfq.edu.ec

Address:

Escuela de Ciencias Biológicas e Ingeniería, Universidad Yachay Tech, Hacienda San José, Proyecto Yachay Urcuquí, Ecuador

Universidad San Francisco de Quito (USFQ), Grupo de Medicina Molecular y Traslacional (MeM&T), Colegio de Ciencias de la Salud (COCSA), Escuela de Medicina, Edificio de Especialidades Médicas, Av. Interoceánica Km 12 ½—Cumbayá, Quito 170157, Ecuador

The present work has been written in

Journal of Chemical Information and Modeling format since the next page

Fishing Potent Tyrosinase Inhibitors with New Molecular Docking Protocol from Natural Products Ocean

Camila Velastegui Gamboa[1], Yovani Marrero-Ponce,[2-3*] Nelson Santiago Vispo[1]

[1]*Universidad de Investigacion de Tecnologia Experimental Yachay, Escuela de Ciencias Biologicas e Ingenieria, San Miguel de Urcuqui, Ecuador*

[2]*Universidad San Francisco de Quito (USFQ), Grupo de Medicina Molecular y Traslacional (MeM&T), Colegio de Ciencias de la Salud (COCSA), Escuela de Medicina, Edificio de Especialidades Médicas, Av. Interoceánica Km 12 ½—Cumbayá, Quito 170157, Ecuador & Computer-Aided Molecular "Biosilico" Discovery and Bioinformatics Research International Network (CAMD-BIR IN), Cumbayá, Quito, Ecuador.*

[3]*Unidad de Investigación de Diseño de Fármacos y Conectividad Molecular, Departamento de Química Física, Facultad de Farmacia, Universitat de València, Spain.*

[*]***Corresponding author*** (✉):

Y. Marrero-Ponce: ymarrero@usfq.edu.ec or ymarrero77@yahoo.es; **Tel.:** +593-2-297-1700 (ext. 4021). http://www.uv.es/yoma/ or http://ymponce.googlepages.com/home**;**

ORCID ID: http://www.orcid.org/0000-0003-2721-1142.

ABSTRACT

Melanin has an important biological role, although their overproduction and accumulation are associated with some diseases going from skin problems as melasma to Parkinson's disease. The overproduction of melanin is due to an increase in the number of melanocytes or melanogenic enzyme activity. The melanocytes are regulated mainly by the tyrosinase enzyme, which catalyzes the two most important steps to produce pigments. This makes the oxidoreductase the principal target to inhibit melanin production. Currently, there are some tyrosinase inhibitors (TIs) used for the treatment of various skin disorders.

Nevertheless, the current TIs had associated some safety problems such as their high cytotoxicity, mutagenesis, non-lasting effects, among others. Therefore, the need for new, more effective, broad-spectrum, safer, and longer-lasting protection of TIs is clear. The existing *"in vitro*" methods to find new inhibitors are expensive and with limited efficacy. The present study will focus on developing a theoretical study with a molecular docking protocol that uses a diverse panel of tyrosinases for the first time and allows the discovery of new TIs using an expert system. Here, we developed classification models with machine learning techniques based on the tyrosinase panel's molecular docking. We developed and integrated all this workflow into free available software SiLis PREENZA for the prediction of TIs.

Furthermore, we made an update of the known tyrosinase inhibitors of 701 to 2514 reported so far. This report will help enhance the process of finding tyrosinase inhibitors to seek novel depigmenting agents for the treatment of hyperpigmentation disorders. The compounds identified as inhibitors could be used to narrow "*in vitro"* experimentations first with phenotypic experiments until they reach melanoma cells.

Keywords: Tyrosinase Inhibitor, Molecular Docking, Autodock-Vina, Machine Learning, Predictive Model, WEKA, SiLis-PREENZA Software, Virtual Screening, Natural Product

Running Title: *Fishing Potent Tyrosinase Inhibitors..*

Graphical Abstract

GLOSSARY

UV: ultraviolet radiation

ROS: reactive oxygen species

CNS: Central Nervous Systems

TIs: Tyrosine inhibitors

QSAR: Quantitative Structure-Activity Relationship

TRP: tyrosinase-related proteins

SM: Supplemental Material

RMSD: Root means square deviation

CA: Cluster analysis

IC_{50}: Half inhibitory concentration

D[1-14]: Data used for developing predictive models

WEKA: Waikato Environment for Knowledge Analysis

k_i: Inhibition constant

ML: Machine Learning

SMOreg: Support Vector Machine for Regression

IBK: K-nearest neighbors' classifier

PSO: Particle Swarm Optimization Search

CC: Correlation coefficient

MAE: Mean absolute error

FLDA: Fisher's Linear Discriminant function

SMO: Support vector classifier

1R: Discretizing numeric attributes

Q: Accuracy

MCC: Matthews Correlation coefficient

ROC: Receiver Operating Characteristic

PRC: Precision-Recall Curve

FP: False Positive

TP: True positive

SiLis-PREENZA: DRY- in silico - Screening & Prediction of Effects of Inhibitors on Enzyme Activity

1. INTRODUCTION

Melanins are pigments present in several organs such as hair, eyes, ears, brain, and skin[1]. This polymer is produced in the melanocytes[2] by a process known as melanogenesis,[3] which includes the synthesis and distribution of the pigment in the epidermis[4]. Melanin plays a crucial role in absorbing light to protect skin against damaging DNA effects of ultraviolet radiation (UV), absorbing free radicals, protect against reactive oxygen species (ROS)[5], and alters the synthesis of vitamin D3[6]. Although melanin's excessive production is associated with several hyperpigmentation problems[7] like melasma, freckles, senile lentigines[5], over-tanning, age spots[1], pigmented acne scars[8], ephelides [9], and so on. Some of the more relevant causes of melanin production are the exposition to the sun, which stimulates epidermal melanocytes, or drug-induced such as minocycline, amiodarone, oral contraceptives, and anticancer drugs or post-inflammatory conditions like a side-effect of laser treatment[4]. Other factors that also induce melanogenesis are alpha-melanocyte-stimulating hormone, melanocortin 1 receptor, and agouti-related protein[10].

The principal focus to treat these problems is tyrosinase (polyphenol oxidase), a glycosylated oxidative enzyme that catalyzes two crucial steps in melanin biosynthesis. First, the hydroxylation of a monophenol like *L*-tyrosine into *L*-DOPA, a catechol (monophenolase activity) and the second the *o*-oxidation of catechols *L*-DOPA into *o*-dopaquinone (diphenolase activity), which is further oxidized to form melanin[8]. Tyrosinase has a binuclear active site with two copper ions and one oxygen molecule,[11] in which the copper ions are coordinated by six histidines residudes[12] and play a crucial role in the activity of the enzyme. Tyrosinase is widely spread in yeast [13], bacteria, fungi, plants, and mammals, especially humans[5]. Therefore, the tyrosinase inhibitors are highly searched for skin-whitening, anti-browning in fruits, and even

used to treat melanomas because they act as an adjuvant [5]. Furthermore, it was discovered that the over-production of tyrosinase in the Central Nervous Systems (CNS) leads to increased dopaquinone levels resulting in neural damage significantly contributing to Parkinson's disease (PD)[3].

Nowadays, several compounds are reported as tyrosinase inhibitors (TIs). For instance, hydroquinone is one of the most famous standard skin-whitening[6]. Although, it has several safety issues due to it could generate ROS and permanently damage melanocytes. The reason why it has been banned from the European Union and is just recommended under dermatologic prescription[14,15]. Another well-known tyrosinase inhibitor is kojic acid, a fungal metabolite, which is more widely used in cosmetics; however, it also has side-effects associated with it dermatitis, sensitization, and carcinogenicity [14]. Some other tyrosinase inhibitors have been identified; nevertheless, most of them had problems related to toxicity[16], low activity, poor target selectivity[3], dermatitis, irritation, leukoderma, hypochromic, ochronosis, between others[2]. Nevertheless, many of them are used as the baseline for the discovery of new tyrosinase inhibitors.

The necessity of finding new potent tyrosinase inhibitors with no side-effects and permanent results is precise. The clinical and industrial demand for TIs is increasing, and one key step in finding new TIs is the virtual screening[16] that will help narrow the bioactive compounds, and the cost of "*in vitro"* experimentation will be lower. Several literature reports used pharmacophore models [7,17,18], Quantitative Structure Activity Relationship (QSAR) [9,19–21], and molecular docking[22–24] in the virtual screening process. Molecular docking shows excellent potential for predicting the action mode in the interaction and binding sites of molecules that are important in drug discovery and therefore, it decreases the experimentation cost and allows the

study of interactions at a molecular level[25] Nonetheless, many docking studies use one or few tyrosinases to develop the predictive models. None of them uses human tyrosinase since there is not currently an X-ray diffraction model of it. Some uses models of human tyrosinase, but any of them uses tyrosinase-related proteins (TRP).

Some of the present authors of this report worked previously with predictive models for TIs using QSAR models. They use data of 701 active compounds reported with tyrosinase activity. This is the most extensive data reported so far, and that has a broad application domain; however, that work was done in 2011[26,27]. Since then, several TIs had been reported, and other predictive models had been made, although any study had included big data that used a broad application domain. Most of the models use small data, congeneric, related to family compounds and just a few models include an application domain. Some reports like Pillaiyar 2018 did a review of the TIs reported, including 141 molecules classified by families[28], even though the purpose of that work is informative and not have been used for predictive purposes. It is necessary to increase the applicability domain from previous works that include the new TIs reported in the last 10 years and update the active TIs.…

The present study will develop a molecular docking protocol that uses a diverse panel of tyrosinases that allows the discovery of new TIs using an expert system. We aim to use software to search new TIs of the ocean of natural products because most of the TIs reported with the highest activity are from this origin. For this purpose, we used molecular docking to build models to predict the inhibitory activity. Here we will use a broad panel of tyrosinase proteins, tyrosinase-related proteins, and human tyrosinase model. This panel of tyrosinases allows us to study the diversity and make a generic screening process with Autodock-Vina. We propose software that helps make the docking with the panel and qualitative and quantitative predictive

models based on docking and facilitating the clustering process. This software is used to find "*in silico"* new TIs. Furthermore, this study made an update of the known TIs, considering the last 10 years.

2. MATERIALS AND METHODS

2.1 Protein Structure Preparation

The seventy-seven three-dimensional (3D) available structures of tyrosinase and tyrosinase-related protein (TRP) structures of X-ray crystal diffraction from ten species (four from bacteria, two from fungus, three from plants, and one from humans) were retrieved from RCSB protein database (http://www.rcsb.org/pdb/home/), and one homologous model of human tyrosinase was obtained from the Protein Model Database[5] (http://srv00.recas.ba.infn.it/PMDB/). A total of nineteen of these proteins, which contain crystal ligand structures, were selected for re-docking studies. AutoDock Tools 1.5.6 software was used for preparing the proteins for docking. This process consisted of removing water molecules and other ligands, adding the polar H-atoms, and converting the 3D structures from PDB to PDBQT format[29].

2.2 Ligand Structure Optimization

Structures were drawn with the program MarvinSketch 19.22[30], the addition of H-atoms and generation of 3D coordinates of the structures were made with ToMoCoMD-CARDD[31] program using 3D-RDkit by using a MMFF94 force field[32]. The conversion from mol to PDBQT was made with Open Babel GUI 2.4.1[33]

2.3 Docking Parameters

To establish the best docking parameters, a set of experiments were carried out. The first experiment aims to determine the use of ions and water in the active site because both play an

important role in the activity of the enzyme[34]. For doing this, we docked 3 proteins that had the same kojic acid ligand from two different species *Bacillus megaterium* (PDB: 3NQ1, 5I38) and *Homo sapiens* (PDB 5M8M). So, we created four variants with the tyrosinase proteins (with water molecules, without water molecules, with ions, and without ions) and compared which variant gets a better fit in the active site to determine in which conditions the docking was the most appropriate. The details of the comparison are on Supplemental Material (SM) at SM1_A.

The second exploration was done to determine the exhaustiveness, which controls how comprehensive is the search to find the best pose for the ligand. The higher the exhaustiveness the results are conclusive, and also more time is invested and therefore for docking a huge amount of data the computational expense will be high[35]. The time increases linearly with the exhaustivity and the probability of finding the global minimum as well. Although the search is from the same seed, large exhaustiveness does not guarantee the global minimum, so it is preferred to take different runs with different seeds that increase the likelihood of finding the pose with minimum energy. The computational cost of an exhaustive search is high due to the time invested in each calculation and in this case finding the best pose[36].

We want to find the best exhaustiveness in which the results are conclusive and do not require much computational expense for docking data libraries. So, we took twelve proteins from two species with different ligands *Bacillus megaterium* (PDB: 4P6R, 4P6S, 4P6T, 5I3A, 5I38, 50AE, 6EI4, 6QXD) and *Homo sapiens* (PDB: 5M8M, 5M8N, 5M8O, 5M8P) and we varied the exhaustiveness for 30, 50, 90 and 100. We examined the minimum exhaustiveness at which the docking has consensus results for the tree runs carried, in other words, the exhaustiveness at which the three runs find the best pose. The results of this experiment are given as supplementary material at SM1_B.

2.4 Self-Docking Studies

In this study, we measured the docked poses' ability to resemble the crystallographic orientations of the ligands. This experiment took 19 tyrosinase proteins that were reported with their ligands from 3 different species, one from *Agaricus bisporus* (PDB: 2Y9X), ten from *Bacillus megaterium* (PDB: 3NQ1,4P6R, 4P6S, 4P6T, 5I3A, 5I3B, 5I38, 50AE, 6EI4, 6QXD) and eight from *Homo sapiens* (PDB: 3KAN, 5M8M, 5M8N, 5M8O, 5M8P, 5M8Q, 5M8R, 5M8T). Each protein was docked with the 11 different ligands present in the crystallization of the different tyrosinases. The root means square deviation (RMSD) of the poses obtained from the docking and the crystallized structure were calculated using the LigRMSD 1.0, a free web-server that calculates RMSD among identical or similar compounds using as reference the crystallized ligand. The strict distance was considered to compare the crystal and dock poses that consider matching identical atoms and bonds[37]. The results of all the calculations are present in SM2.

2.5 Selection of tyrosinase panel

A set of 42 PDBs were selected, and one extra homology model of human tyrosinase was found in the Protein Model Database (see SM3_A for a detailed description of the PDBs). From this set, several proteins were selected based on the diversity of ligands and species, aiming to provide a wide range of tyrosinase protein representations. To reduce this panel of tyrosinase proteins, a cluster analysis (CA) was carried. The CA helps make a profound and complementary comparison of the tyrosinase proteins; a tree-based CA was used. Here, we used the docking affinities calculated with Autodock-Vina for 88 ligands identified as strong, intermediate, and weak TIs based on experimental data with a reported half inhibitory concentration (IC_{50}) see SM3_B, which indicate how much of the drug is needed to inhibit tyrosinase activity by half[38].

These 88 ligands were docked according to the docking parameters established in the previous section against the proteins selected. We performed in STATISTICA software hierarchical agglomerative clustering using Ward's method and the squared Euclidean distance as an amalgamation rule and proximity function, respectively[39]. The description of the final panel of 13 proteins discovered is in Table 1, and the PDBQT and the configuration files are given as SM3_C.

Table 1 comes about here

2.6 Predictive Models

Fourteen data sets (denoted as **D1-D14**) from 15 reports were selected to build predictive models and calibrate the scoring function for regression and classification using Waikato Environment for Knowledge Analysis (WEKA) software 3.9.4 [40]. Data from one to twelve were used in their papers for building predictive models or did a molecular docking study. These data had less than 100 compounds and the majority of them focused on one specific family of compounds[7,9,17–22,24,41–43]. Data thirteen, **D13**, is an actual literature review of the 144 TIs (91 with reported activity). This review explores various family compounds, so the data is varied and their application domain[28]. Finally, **D14** was extracted by QSAR models of two papers of one of the authors of this study[26,27]. This is a compilation of 1422 molecules of the two papers with 701 active TIs. This is the most extensive data reported so far because even tyrosinase reviews just reported around 150 molecules [16,44–47] and some reported predictive models were built with less than 100 compounds like the ones of **D1-D12** used here, further details about data are in Table 2. **D14** also has a broad application domain due to the diversity and quantity of molecules present.

Table 2 comes about here

Finally, we also elaborated a final data set **D15,** a recompilation of the reported tyrosine inhibitors from around 185 articles of the last 10 years, and it was used as an external validation

to assess the models. It is essential to highlight that **D15** does not have any common molecules with **D14**. Although, it contains all the molecules from the data **D1-D13**. The compilation of D14 and D15 is available on SM4_D15, and it can be further used for future studies on the field of TIs. A detailed description of all the data previously described is in Table 2.

First, every molecule was docked against the panel of 13 proteins selected (see Table 1) in the previous step. From the three-run docked procedure 4 parameters were calculated for every molecule and protein, which are: 1) the average of the nine best affinity positions 2) the maximum affinity, 3) the minimum affinity, and 4) leader efficiency, which is the average affinity divided by the number of heavy atoms. All these parameters were used as the independent variables for the training of the predictive models.

2.6.1 Regression Models

For the regression models, the variables were the 4 docking parameters of the 13 proteins and the activity like IC_{50}, inhibition constant (k_i), or % of inhibition, see SM4. The regression models were built to predict the values of activity, and based on the numeric prediction, a class was assigned according to some cut-offs (see SM4-BP for a detailed description of the breakpoints to each data set).

First, the wrapper selection of WEKA attributes was applied to obtain subsets of different variables that can predict the activity. The wrapper selection evaluates attribute sets by using a learning scheme[48]. Six different Machine Learning (ML) classifiers (Gaussian Processes, Linear Regression, Support Vector Machine for Regression (SMOreg), K-nearest neighbors' classifier (IBK), M5' model tree algorithm and Random Forest) and 3 different Search Method (Genetic Search, Greedy Stepwise, and Particle Swarm Optimization Search (PSO)) were used to form the subsets. The configurations of the ML techniques were set as default in WEKA software. From

this procedure, we obtained all the possible subsets to evaluate the regression. The training assessment was performed with Cross-Validation 10, Split 66, and All training set.

The prediction ability was evaluated by considering the correlation coefficient (CC) and mean absolute error (MAE). In the case of being necessary it was applied meta-classifiers to improve the scores of the models. We used Additive Regression, Bagging, Stacking, and Voting varying for the meta-classifiers and using the same six ML techniques used for building the models. The best two models for each data set for regression were selected for further evaluation. The SDF file, model, detailed description, and statistics of the models selected can be found in SM4.

2.6.2 Classification Models

The classification models were done with the same independent variables as regression models. In general, compounds with an IC_{50} less than 10 uM or ki less than 4.7 were considered active and the rest inactive (see SM4-BP). Then, we made the attribute selection with 7 classifiers (Naïve Bayes, Fisher's Linear Discriminant function (FLDA), Logistic, John Platt's sequential minimal optimization algorithm for training a support vector classifier (SMO), K-nearest neighbors' classifier (IBK), The minimum-error attribute for prediction, discretizing numeric attributes(1R), and Random Forest) and the same 3 search methods used for regression models.

We obtained all the different subsets, and we evaluate performance in classification based on the 3-test options mentioned in the former section. The main statistical parameters taking into consideration were Accuracy (Q), Matthews Correlation coefficient (MCC), area under the Receiver Operating Characteristic (ROC) curve, Precision-Recall Curve (PRC), Precision, False Positive (FP) rate, and True positive (TP) rate statistics[49,50]. After selecting the best subsets, the

meta classifiers were applied to improve the performance of the models. The meta-classifiers used were AdaBoost M1, Bagging, Stacking, and Voting, with the rules of combination of each case. Then the best two models were selected for each data set for further evaluation (see SM4).

2.6.3 External Validation

Finally, the best models for regression and classification of every data set were tested against external data of 1813 compounds (**D15**) to provide further assessment of the robustness and the predictive power of the models' performance. These 1813 molecules were docked and then validated against the models considering three different activity breakpoints 1, 10, and 30 uM. The cutoff of 1 was to build a more restrictive model and, therefore, can accurately discriminate between active and inactive. The breakpoint of 10 is because most of the models were built with that threshold for activity, and the breakpoint of 30 was chosen because it was the breakpoint activity of **D14**, which is the largest data.

2.6.4 Consensus Models

To improve the individual models' predictive power, several consensus models were developed to enhance the classification. Consensus approaches aim to combine and fuse the outcome from different sources to increase the outcome reliability compared to individual models[51]. Instead of finding the best model, the aim was to find the best subset of individual classifiers that predict the external data **D15**. To find the best consensus models, several combinations of the best predictive performance models were done, taking into consideration de accuracy (Q), kappa statistic, and F1 score. The combinations were done for all three breakpoints and were evaluated for all of them. The combination rule was made manually by the majority voting process without supervision. The best two consensus models were selected for each activity breakpoint (denoted as **D15_P01A, D15_P01B, D15_P10A, D15P_10B, D15_P30A,**

D15_P30B). These consensus models improve the outcome of predictions of the base models. The statistics of all the consensus models generated for each breakpoint are available on SM5.

2.7 Software Development

We developed a software freely available named SiLis-PREENZA, the acronym for DRY- in silico - Screening & Prediction of Effects of Inhibitors on Enzyme Activity all the process previously described since docking until consensus predictive models are implemented. Therefore, the process of finding TIs is automated. This software was implemented in Java 1.8 programing language; hence, it is a cross-platform software. For developing the SiLis PREENZA software Open Babel 3.0[33], RDKit[52] AutoDock Vina[53], WEKA[40], 2D QuBiLS-MAS[54], QuBiLs-MIDAS[31] toolkits were used for the manipulation of structures, docking, predictive models, and clustering.

2.8 Prospective Virtual Screening

This software was used for the prospective screening in which we use eight data sets for screening and the discovery of new TIs. We selected data that was identified as natural products since the main known TIs are from natural origin like kojic acid discovered from a fungus *Aspergillus flavus*[55], arbutin a phenolic compound found in some plants[56,57], ascorbic acid a vitamin commonly found in fruits[58], mimosine a natural amino acid found in plants[59] among others. These data sets were denominated as **S1-S8**. The screening data summary is in Table 3, and the SDF of all the data is on SM6.

Table 3 comes about here

Finally, all the molecules were passed through all the predictive models used for the consensus models, and the selection of possible tyrosinase inhibitors was carried considering 2 rules. The first rule is that the molecules should be active for the two consensus models for 1 uM breakpoint, or it could be active for one model of 1 but active for both models of 10uM.

2.8.1 Cluster Analysis

After selecting the active molecules for the consensus models, we carried out cluster analysis, also implemented on SiLis-PREENZA software, of the virtual hits to suggest the diverse ones for experimentation. The cluster analysis was made using the k-means and hierarchical methods that take variables 50 descriptors, 25 topological 2D QuBiLS-MAS[54] and 25 Geometrical 3D QuBiLS-MIDAS[31] descriptors. The number of clusters selected was 20, and we searched for similar molecules and have the same chemical core to just select one representative.

After this, we carried out a second cluster analysis with the virtual hits selected to form the natural products data and all the reference compounds reported as tyrosinase inhibitors that are the 701 positive inhibitors of **D14** and all the 1813 compounds **D15** a total of 2514 reported TIs. We carried out the cluster analysis with Che-S Mapper software[60] with the k-cascade algorithm with CDK descriptors, and we selected 30-60 clusters. Then we compared the molecules to the reported ones to select the molecules that represent new cores. The Table of the detailed selection procedure is available on SM7.

The final corroboration was done by docking again the selected hits by 3 repetitions and selecting just the ones that were active at least by two docking runs of the models of 1uM. From this selection we search for the availability of compounds in PubChem database. Finally, we selected some of the best virtual hits and compared them with the docking of the 4 known TIs arbutin, hydroquinone, kojic acid, and mimosine. We compared their binding affinities and the patron of residues interactions using Protein-Ligand Interaction Profiler that detects and visualize interactions among protein and ligands[61]

3. RESULTS AND DISCUSSION

3.1 Docking Parameters

The conditions of docking parameters were established based on the experiments described in the material and methods section. All the docking runs were performed with Autodock-Vina 4.2.6[53]. We obtained that conserving the metal ions and removing water molecules suit the best ligands from the first experiment. From the second experiment, we obtained that the best exhaustiveness is 30. Therefore, we conclude that the binding site comprises two copper ions or two zinc ions depending on the crystal structure. The docking site for the ligands on tyrosinase structures is defined by establishing a cube at the protein's active site in the geometric center of the co-crystalized ligand, with dimensions of 40x40x40 Å, and employing a grid box spacing of 0.375 Å. Three runs are needed for each compound with exhaustiveness equal to 30, energy range 3, and the first 20 positions with the highest values of binding affinity are rescued; see SM1 for more details.

3.2 Self-docking

Molecular Docking validation was done with the self-docking or redocking, which was done comparing the ligands' crystal pose and the poses docked. The results revealed that the docked poses resemble great the experimental ones. For instance, in the PDB 5M8M the redocking of the kojic acid and the crystalized kojic acid have an RMSD of 0.25 (see Figure 1). The RMSD obtained to support the docking protocol's accuracy and even it is better than other reports that their RMSD for kojic acid in the PDB 3NQ1 is 1.36[22].

Figure 1 comes about here

For illustrative purposes, the best poses obtained from the main inhibitors from both *Bacillus megaterium* and *Homo sapiens* are shown in Figure 2. These results showed the

reliability of using the docking protocol to reproduce the ligands' experimental conformations in the enzyme tyrosinase or TRP.

Figure 2 comes about here

3.3 Discovering the Tyrosinase Panel

Initially, several tyrosinase proteins were proposed 47 (see SM3_A). So, we first filter tyrosinases considering the resolution, diversity, crystallization, resolution, and 29 proteins. All these proteins were docked against the 11 ligands found crystallized in PDBs, and then the cluster analysis allows us to select just 17 proteins. To decrease the panel, even more, 88 ligands classified as strong, intermediate, and weak were docked with Autodock-Vina, and the average of the best nine docking positions was used to make a cluster analysis (see SM3_B). The results of the CA using Ward's methods are shown in a dendrogram of Figure 3.

Figure 3 comes about here

We analyzed just the clades that directly related two leaves for selecting the proteins based on the CA and their dendrogram. The first clade with two leaves we have 5OAE and 5I38, both proteins are from *Bacillus megaterium* and have ligands. Although 5I38 had better resolution than 50AE and their ligand is a recognized TI, so we selected this protein. In the second clade with two leaves, we had 5M8R and 3KAN both are from *Homo sapiens***,** but 3KAN is the only protein of TRP2, so it was selected. We found 5M8O and 5M8M in the third clade, and we selected 5M8M based on the Euclidean distances. We found the four clades the proteins 6HQI and 5Z0D; both were from different species, but 5Z0D has better resolution so it was chosen. The last paired clade had the proteins PM0079416 and 6ELS, the first one is the model of human tyrosinase, and the second is from *Malus domestica* (apple). In this case, both were

representative and diverse; we selected both proteins. Finally, the proteins that were not paired in a clade were directly selected, such as 6QXD, 3NQ1, 5M8Q, 5ZRD, 3W6W, 5CE9, and 2Y9X.

Thirteen proteins were selected to constitute the tyrosinase panel considering their representability from the point of view of the structure, affinity with docking, and origin. The PDBs ID and species of these proteins are 2Y9X (*Agaricus bisporus*), 3KAN (*Homo sapiens)*, 3NQ1 (*Bacillus megaterium*), 3W6W (*Asperugillus oryzae*), 5CE9 (*Juglans regia*), 5I38 (*Bacillus megaterium*), 5M8M (*Homo sapiens)*, 5M8Q (*Homo sapiens)*, 5Z0D (*Streptomyces castaneoglobisporus*), 5ZRD (*Burkholderia thailandensis*), 6ELS (*Malus domestica)*, 6QXD (*Bacillus megaterium*), and PM0079416 (*Homo sapiens*. A detailed description of the PDBs selected is in Table 1, and the PDBQT was deposited with the configuration file in SM3_C. This is the first time that someone uses a broad and diverse panel of tyrosinases to make docking that includes tyrosinases from several species, tyrosinase-related proteins, and a model of human tyrosinase.

3.4 Scoring Function Calibration & Predictive Models

Classification and Regression models aim to predict if a compound is a tyrosinase inhibitor (ACTIVE) or not (INACTIVE). For doing this, we use a data set with 52 independent variables (4 derived affinity docking variables of the 13 proteins) plus their respective response variable (e.g., IC_{50}) to build the models. The Wrapper function of WEKA was first used to generating small subsets. Models of data from 1 to 13 were developed with compounds reported with some tyrosinase activity, and Models of D14 with compounds reported as active in the literature and inactive as never tested against tyrosinase. Approximately 50 models of classification and regression were selected with the best predictive power for 14 data sets

described in the former section. The summary of each regression model and classification and their performance with their data is in Table 4.

Table 4 comes about here

The predictive models presented here had in general better performances than the models reported in the original papers where the molecules come from. For instance, the D1_RA model had a square correlation coefficient (CC) of 0.8898 for cross-validation 10, and the original model reported a CC of 0.745[7] for the test set, it is important to note that although both statistics are not the same, both are the type of evaluation which leaves one group of molecules to validate the model and the rest to build it. Just model of D8 has similar statistics like the ones reported here, and in that case, they reported a CC for cross-validation 10 of 0.74 with 13 outliers[20] and our model has a CC of 0.7428 but with just 3 outliers considered. These values indicate that our model has a higher CC and even taken into consideration a high variation of the molecules. The table with a detailed comparison of our models and the original is available on SM4_C.

3.4.1 External Evaluation

To evaluate the real power of prediction of the models developed it was probed against external data which we denominate **D15** that has 1813 compounds and represent the known space of tyrosinase inhibitors of the last 10 years without taking into consideration all the 701 active molecules present in **D14**. The top results of the evaluation of these models by the 3 activity breakpoints are in Table 5.

Table 5 comes about here

The model that best predicts for 1 as the activity breakpoint is **D10RB** with an accuracy of 0.846 and then **D3RA** with an accuracy of 0.837. Although the accuracy is relatively high, the kappa statistic is very low; for **D10RA** the model predicts all the molecules as inactive, and like

the **D15** with breakpoint 1 has few actives, it has a higher score. Based on kappa statistic **D7CB** with an accuracy of 0.71 and **D13CB** with an accuracy of 0.547, the best ranked and considering the F1 active score **D8CA** with 0.413 and **D13CB** are the best ranked. It means that these models are not good at recovering the active molecules.

For the breakpoint of 10, the top models considering accuracy are **D10RB** with 0.629 and **D3RA** with 0.628. Nevertheless, it happens the same problem for the activity breakpoint of 1, and the models best ranked for kappa are **D7CB** with 0.614 and **D7CA** with 0.58, and considering F1 active score is **D12RA** and **D12CB** with accuracy 0.371 and 0.397 respectively.

Finally, for breakpoint of 30, the best models in accuracy and kappa statistic are **D8CA** and **D7CB** with 0.547 and 0.544 respectively of accuracy. If the active and inactive compounds are balanced, the best models based on the kappa statistic and accuracy are the same. Furthermore, the best models considering F1 active scores are **D12RA** and **D2CB** with 0.494 and 0.514 accuracy. As the results have shown, the individual models do not predict **D15** very well; the highest accuracy is that the models predict most compounds as inactive, and therefore various active compounds are lost. To solve this problem, consensus models were created to improve the models' statistics for the 3 breakpoints.

3.4.2 Consensus models

Several combinations considering the models with the best parameters were done to each breakpoint to improve the performance of the models (see Table 5). There were choose two consensus models for each breakpoint that improve the accuracy, kappa statistic, and F1 active score. The best two models for 1 uM are one of 9 variables denominated **D15_P01A** with an accuracy of 0.782 and **D15_P01B** with an accuracy of 0.697, which relatively has lower accuracy of the best individual models although their statistics are improved. For the breakpoint

of 10, there are 2 models, one or 9 variables **D15_P10A** with an accuracy of 0.65 and **D15_P10A** with 0.601, the model **D15_P10A** had higher accuracy than the best individual model of 10, and their statistics are good also. Finally, for the threshold of 30, the consensus models of 7 and 9 variables (**D15_P30A** and **D15_P30B,** respectively) were selected with 0.549 and 0.591 as accuracy, which is better than the best individual models reported.

3.5 SiLis-PREENZA Software for the discovery of new TIs

The SiLis-PREENZA Software is composed of a friendly desktop user interface (see Figure 4), and the procedure can be summarized in 5 steps: (1) The selection of the data that can be in SDF or MOL files. The data could be in 2D or 3D structures, in the case of 2D the software also can generate the 3D structure in step (2) Generation of 3D Structures with RDKit[52] that uses two different force fields: i) Molecular Mechanic Force Field (MMFF94) or ii) Universal Force Field (UFF). (3) Selection of the predictive models. In this step, we could select the proteins for docking, the predictive models of classification, including **D15** models or regression. Here we can also select the applicability domain and the time out function (set as 10 min by default). (4) Selection of Clustering methods. This step was implemented to evidence the resemblance among molecules with cluster analysis. We can select 3 types: i) Simple expectation maximization, ii) K-means, or iii) Hierarchical. The clustering can be done with 25 topological 2D descriptors or 25 geometrical 3D descriptors, and lastly, step (5) Selection of the folder to save the results.

Figure 4 comes about here

The time-out function was implemented because docking is a time-requiring step, and depending on the complexity of a molecule, it could take days to dock to one protein. The function was implanted for optimizing the docking procedure and the time spent in this step. The docking of one molecule against one protein is estimated based on 5 CDK descriptors nAtom,

nHBDon, TopoPSA, XLogP, nRotB, which corresponds to the number of atoms, number of H-Atoms as Donors, Topological Polar Surface Area, partition coefficient, and the number of Rotatable bonds, respectively. The function is an assemble of 3 ML techniques that estimate the average time. If the estimation is specified over time, then the molecule will not be docked, and time will be saved with overcomplicated molecules. We used 65 molecules of different complexity for modeling this function, and we measure the time invested in docking of different tree proteins. For more details about the time-out function, see SM4-TO. After applying this function the 10% of molecules were not docked in 65 molecules, and for the prospective screening of natural products, 12.5% of molecules were not docked.

After the screening, an interface for the processing of the results is shown, and the folder with all the files and procedures involved is generated where we can rescue any file that we want from docking, predictive models, or clustering. Moreover, **D13, D14,** and **D15** used for modeling in this study are available as sample tests in the first step. The software presented here can be used for screening for tyrosinase inhibitors quickly and automatically and includes many functions like docking predictive models and clustering being the first of its class in the field of tyrosinase inhibitors.

3.6 Prospective Virtual Screening

Natural products are critical in drug discovery because they contain scaffold diversity and structural complexity[62]. So, if we want to find new scaffolds of TIs, natural products are one of the main compound's pools for prospective screening. In addition to that, several TIs are from natural origin, making the Natural products' database ideal to find new TIs. We selected extensive data of natural products for prospective screening employing SiLis-PREENZA Software for all the reasons mentioned before. We used the models **D15_P01A, D15_P01B,**

D15_P10A, and **D15P_10B** to select the virtual hits based on 2 rules: i) active for both models **D15_P01A** and **D15_P01B** or ii) active for one model of 1 and both**, D15_P10A**, and **D15P_10B.**

The screening process results of each data and the virtual hits are presented in Table 3 and the output of the software for the models is on SM6. In general, several active compounds in each data set in average 64 molecules from which the second rule selected 11. However, to find new TIS it is necessary to select just the molecules that represent new scaffolds. For this purpose, we carried out two cluster analyses to maximize the molecules' diversity with the smaller number of molecules.

The first cluster analysis was among the virtual hits of the 8 natural products data sets (512 in total). This CA was done by two hierarchical and k-means using the SiLis PREENZA software (See SM7_A). Based on both CA, we selected 129 molecules markedly different (see Table 3 for more details).

After obtaining the virtual hits that were markedly different among them, we need to compare them with tyrosinase inhibitors' available space. Hence, we carried out a second cluster analysis with 2514 molecules (active of **D14** and **D15**) and 129 virtual hits. This CA was done using CheS Mapper [60] with CDK descriptors. The software selected 30 clusters, and the active virtual hits were present in 22 clusters (See SM). CheS Mapper software arranges compounds in a 3D space, in which the spatial proximity reflects their similarity[60] as can be observed in Figure 5. So, to determine the compounds similar to the virtual hits, visual analysis was carried out.

Figure 5 comes about here

Cluster 1 contains 29 molecules, and 2 of them (7%) are virtual hits. We looked for the closest neighbor of the reference compounds and compared the structures. If the structures are

different, they are selected as a virtual hit, but the virtual hit is not chosen if the structures are similar. For instance, in cluster 1 the first virtual hit was different from the closest neighbor reported with an IC_{50} of 19.20[28] and it was selected as a virtual hit. On the other side, the other virtual hit was the duplicate of Dodecandioic acid (the reference compound), so it was not selected, see Figure 6.

Figure 6 comes about here

The same analysis was employed for all the clusters and virtual hits; see SM7_C for the detailed comparison of all VH with their closet neighbor. After this, the initial data was reduced to 87 compounds predicted with tyrosinase activity and that have a different scaffold from the reported TIs. The prominent representatives by a cluster of the virtual hits are in Figure 7.

Figure 7 comes about here

The final filter selection of compounds was done using SiLis-PREENZA Software, in which we pass again the 87 compounds for all D15 and regression models. D15 models were used to select just the molecules that in the tree runs continue to be active, and the regression models were done to calculate the average IC_{50} to have an idea of the best virtual leads. From these studies we reduce the virtual hits to 57 compounds, in which 8 molecules were always active for all the **D15** models, these compounds were prioritized for comparative studies plus two compounds that were active in the tree repetitions for models of 1 uM and that their average IC50, minimum and maximum for the regression models was the lowest. The final virtual leads were searched in the PubChem database to look if they already are reported as a TIs or other related purposes and many of them did not have any biological test or patent, and the ones that had reported activity was not related to tyrosinase the table with a detailed description is on SM8_A

In general, our fished molecules had higher affinity values than the 4 TIs used for comparative studies. The docking affinity of our 10 lead compounds ranges from -5.2 to -6.93 on average for the model of Human Tyrosinase (see_SM8_B). These docking values indicate stable enzyme-substrate interactions and better docking scores of previous virtual hits found by the QSAR model which range from -5.8169 to -47332[9]. Other studies reported affinity values ranged from -5.50 to -9.81 and all of them had IC_{50} for tyrosinase inhibition reported[17], which suggests that our virtual hits are within the range of TIs and are reliable for further experimentation against tyrosinase.

The fished molecules also had similar interaction patterns with the four know TIs and with the residues found experimentally in crystal structures that are interacting with inhibitors (see SM8_C). For instance, the crystal structure of PDB:2y9x from *Agaricus bisporus* with an inhibitor tropolone that has interactions with 3 key residues 283VAL, 263HIS, and 264PHE[63], these residues are interacting as well with the know TIs and in the case of prioritizing virtual leads had interactions with the same residues. In fact, Liridine known as compound 80 has hydrophobic interactions with 228VAL, 264 PHE, 283VAL, and hydrogen bond with 244HIS (see Figure 8). In the case of PDB:3nq1, a tyrosinase from *Bacillus megaterium* with kojic acid, there are reported interactions with 197PHE, 201PRO, 205ASN, and 209ARG[64] and compound 80 has hydrophobic interactions with 197PHE, 201PRO, 205ASN, hydrogen bonds with 209ARG, 218VAL and π-Stacking with 209ARG. As we can see all the residues reported experimentally are interacting with compound 80.

Figure 8 comes about here

In the case of tyrosinase of *Homo sapiens* the crystal structure of a TRP1 (PDB:5m8m) is also crystallized with kojic acid and has interactions with 381HIS, 374ARG, and 394SER[65], and

Compound 80 present hydrophobic interactions with 378ASN, 381HIS, and 382LEU, hydrogen bonds with 362TYR and 374ARG which show an alike pattern of interactions. Finally, the model of human tyrosinase does not have a ligand, so we compared the interactions with 4 known TIs. Liridine had hydrophobic interactions with 85HIS and 260VAL. In the case of the 4 TIs used for comparative purposes here, all have hydrophobic interactions with 260VAL, and just arbutin and mimosine interacted with 85HIS. The same analysis was done for the other fished hits and using the 13 proteins of the panel (see SM8_C), and we found that there are similar interactions among them and the 4 TIs, which suggests that our fished hits are suitable candidates to proceed to experimentation and probe tyrosinase inhibitory properties.

4. CONCLUDING REMARKS

TIs are highly searched because they can be used in the cosmetic, food, and medical industry due to their presence in several organisms. The search for new tyrosinase inhibitors is evident because, so far, any TIs are safe, with long-lasting results, and potent. One of the focal points in this search was the natural products due to many reported TIs are from this origin. In this context "in silico" approaches are essential because it reduces the financial cost and the time of the exploration.

In this report, we used molecular docking with a broad panel of tyrosinases and machine learning techniques to build predictive models capable of identity TIs. We developed an expert system named SiLis-PREENZA that internally implements the workflow of different toolkits that employing models based on molecular docking allow us to predict the activity against the tyrosinase enzyme. The developed models were used for prospective screening of natural products and were able to identify possible TIs that are diverse in structure from the previous

reported TIs and therefore are worthwhile to the made inhibitory assay with them. Here, we also made an update of the know tyrosinase inhibitors, so there is an increase in the applicability domain. The outcomes developed in this study can be further use to find new tyrosinase inhibitors.

5. FUTURE OUTLOOKS

Regarding the good behavior that this computational approach brings through SiLis-PREENZA, we seek to make inhibitory essays of the molecules selected by the models like TIs, which are the most differences from reference molecules find, and in the ultimate instance that are cheap. Furthermore, we will implement QSAR models in the software as well to predict TIs

Supplementary Materials:

The Supporting Information is available free of charge on

6. REFERENCES

1. Lai, X., Wichers, H. J., Soler-Lopez, M. & Dijkstra, B. W. Structure and Function of Human Tyrosinase and Tyrosinase-Related Proteins. *Chemistry - A European Journal* **24**, 47–55 (2018).

2. Feng, L. *et al.* De Novo Molecular Design of a Novel Octapeptide That Inhibits in Vivo Melanogenesis and Has Great Transdermal Ability. *Journal of Medicinal Chemistry* **61**, 6846–6857 (2018).

3. Li, Q. *et al.* Identification by shape-based virtual screening and evaluation of new tyrosinase inhibitors. *PeerJ* **2018**, 1–14 (2018).

4. Tang, H. C. & Chen, Y. C. Identification of tyrosinase inhibitors from traditional Chinese medicines for the management of hyperpigmentation. *SpringerPlus* **4**, (2015).

5. Favre, E., Daina, A., Carrupt, P. A. & Nurisso, A. Modeling the met form of human tyrosinase: A refined and hydrated pocket for antagonist design. *Chemical Biology and Drug Design* **84**, 206–215 (2014).

6. Chen, W. C. *et al.* Discovery of highly potent tyrosinase inhibitor, T1, with significant anti-melanogenesis ability by zebrafish in vivo assay and computational molecular modeling. *Scientific Reports* **5**, 1–8 (2015).

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

8. Hassan, M., Ashraf, Z., Abbas, Q., Raza, H. & Seo, S. Y. Exploration of Novel Human Tyrosinase Inhibitors by Molecular Modeling, Docking and Simulation Studies. *Interdisciplinary Sciences: Computational Life Sciences* **10**, 68–80 (2018).

9. Sawant, R. L., Lanke, P. D. & Wadekar, J. B. Tyrosinase inhibitory activity, 3D QSAR, and molecular docking study of 2,5-disubstituted-1,3,4-oxadiazoles. *Journal of Chemistry* **2013**, (2013).

10. Ashraf, Z. *et al.* Carvacrol derivatives as mushroom tyrosinase inhibitors; synthesis, kinetics mechanism and molecular docking studies. *PLoS ONE* **12**, 1–17 (2017).

11. Oyama, T. *et al.* Structural insight into the active site of mushroom tyrosinase using phenylbenzoic acid derivatives. *Bioorganic and Medicinal Chemistry Letters* **27**, 2868–2872 (2017).

12. Ferro, S. *et al.* Targeting Tyrosinase: Development and Structural Insights of Novel Inhibitors Bearing Arylpiperidine and Arylpiperazine Fragments. *Journal of Medicinal Chemistry* **61**, 3908–3917 (2018).

13. Fairhead, M. & Thöny-Meyer, L. Bacterial tyrosinases: Old enzymes with new relevance to biotechnology. *New Biotechnology* **29**, 183–191 (2012).

14. Deri, B. *et al.* The unravelling of the complex pattern of tyrosinase inhibition. *Scientific Reports* **6**, 1–10 (2016).

15. Mann, T. *et al.* Inhibition of Human Tyrosinase Requires Molecular Motifs Distinctively Different from Mushroom Tyrosinase. *Journal of Investigative Dermatology* **138**, 1601–1608 (2018).

16. Zolghadri, S. *et al.* A comprehensive review on tyrosinase inhibitors. *Journal of Enzyme Inhibition and Medicinal Chemistry* **34**, 279–309 (2019).

17. Sari, S., Barut, B., Özel, A. & Şöhretoğlu, D. Tyrosinase inhibitory effects of Vinca major and its secondary metabolites: Enzyme kinetics and in silico inhibition model of the metabolites validated by pharmacophore modelling. *Bioorganic Chemistry* **92**, 103259 (2019).

18. Hsiao, N. W. *et al.* Serendipitous discovery of short peptides from natural products as tyrosinase inhibitors. *Journal of Chemical Information and Modeling* **54**, 3099–3111 (2014).

19. Xue, C. bin *et al.* 3D-QSAR and molecular docking studies of benzaldehyde thiosemicarbazone, benzaldehyde, benzoic acid, and their derivatives as phenoloxidase inhibitors. *Bioorganic and Medicinal Chemistry* **15**, 2006–2015 (2007).

20. Tang, H., Cui, F., Liu, L. & Li, Y. Predictive models for tyrosinase inhibitors: Challenges from heterogeneous activity data determined by different experimental protocols. *Computational Biology and Chemistry* **73**, 79–84 (2018).

21. Dong, H., Liu, J., Liu, X., Yu, Y. & Cao, S. Molecular docking and QSAR analyses of aromatic heterocycle thiosemicarbazone analogues for finding novel tyrosinase inhibitors. *Bioorganic Chemistry* **75**, 106–117 (2017).

22. Tanguenyongwatana, P. & Jongkon, N. Molecular docking study of tyrosinase inhibitors using ArgusLab 4.0.1: A comparative study. *Thai Journal of Pharmaceutical Sciences* **40**, 21–25 (2016).

23. Radhakrishnan, S., Shimmon, R., Conn, C. & Baker, A. Design, synthesis and biological evaluation of hydroxy substituted amino chalcone compounds for antityrosinase activity in B16 cells. *Bioorganic Chemistry* **62**, 117–123 (2015).

24. Channar, P. A. *et al.* Synthesis, computational studies and enzyme inhibitory kinetics of substituted methyl[2-(4-dimethylamino-benzylidene)-hydrazono)-4-oxo-thiazolidin-5-ylidene]acetates as mushroom tyrosinase inhibitors. *Bioorganic and Medicinal Chemistry* **25**, 5929–5938 (2017).

25. Tao, X. *et al.* Recent developments in molecular docking technology applied in food science: a review. *International Journal of Food Science and Technology* **55**, 33–45 (2020).

26. Le-Thi-thu, H. *et al.* Novel coumarin-based tyrosinase inhibitors discovered by OECD principles-validated QSAR approach from an enlarged, balanced database. *Molecular Diversity* **15**, (2011).

27. Le-Thi-Thu, H. *et al.* A comparative study of nonlinear machine learning for the “in silico” depiction of tyrosinase inhibitory activity from molecular structure. *Molecular Informatics* **30**, (2011).

28. Pillaiyar, T., Namasivayam, V., Manickam, M. & Jung, S. H. Inhibitors of Melanogenesis: An Updated Review. *Journal of Medicinal Chemistry* **61**, 7395–7418 (2018).

29. Forli, S. *et al.* Computational protein-ligand docking and virtual drug screening with the AutoDock suite. *Nature Protocols* **11**, (2016).

30. Volford, A. MarvinSketch. *Chem Axon MarvinSketch* (2015).

31. García-Jacas, C. R. *et al.* QuBiLS-MIDAS: A parallel free-software for molecular descriptors computation based on multilinear algebraic maps. *Journal of Computational Chemistry* **35**, (2014).

32. Casañola-Martín, G. M. *et al.* TOMOCOMD-CARDD descriptors-based virtual screening of tyrosinase inhibitors: Evaluation of different classification model combinations using bond-based linear indices. *Bioorganic and Medicinal Chemistry* **15**, (2007).

33. O'Boyle, N. M. *et al.* Open Babel: An Open chemical toolbox. *Journal of Cheminformatics* **3**, (2011).

34. Zou, C. *et al.* Determination of the bridging ligand in the active site of tyrosinase. *Molecules* **22**, 1–10 (2017).

35. Hassan, N. M., Alhossary, A. A., Mu, Y. & Kwoh, C. K. Protein-Ligand Blind Docking Using QuickVina-W with Inter-Process Spatio-Temporal Integration. *Scientific Reports* **7**, (2017).

36. Jaghoori, M. M., Bleijlevens, B. & Olabarriaga, S. D. 1001 Ways to run AutoDock Vina for virtual screening. *Journal of Computer-Aided Molecular Design* **30**, (2016).

37. Velázquez-Libera, J. L. *et al.* LigRMSD: A web server for automatic structure matching and RMSD calculations among identical and similar compounds in protein-ligand docking. *Bioinformatics* **36**, (2020).

38. Aykul, S. & Martinez-Hackert, E. Determination of half-maximal inhibitory concentration using biosensor-based protein interaction analysis. *Analytical Biochemistry* **508**, (2016).

39. Rivera-Borroto, O. M., Marrero-Ponce, Y., García-De La Vega, J. M. & Grau-Ábalo, R. D. C. Comparison of combinatorial clustering methods on pharmacological data sets represented by machine learning-selected real molecular descriptors. *Journal of Chemical Information and Modeling* **51**, (2011).

40. The WEKA workbench. in *Data Mining* (2017). doi:10.1016/b978-0-12-804291-5.00024-6.

41. Ghayas, S. *et al.* 3D QSAR pharmacophore-based virtual screening for the identification of potential inhibitors of tyrosinase. *Journal of Biomolecular Structure and Dynamics* **38**, 2916–2927 (2020).

42. Choi, J., Choi, K. E., Park, S. J., Kim, S. Y. & Jee, J. G. Ensemble-Based Virtual Screening Led to the Discovery of New Classes of Potent Tyrosinase Inhibitors. *Journal of Chemical Information and Modeling* **56**, 354–367 (2016).

43. Radhakrishnan, S., Shimmon, R., Conn, C. & Baker, A. Integrated kinetic studies and computational analysis on naphthyl chalcones as mushroom tyrosinase inhibitors. *Bioorganic and Medicinal Chemistry Letters* **25**, 4085–4091 (2015).

44. Loizzo, M. R., Tundis, R. & Menichini, F. Natural and Synthetic Tyrosinase Inhibitors as Antibrowning Agents: An Update. *Comprehensive Reviews in Food Science and Food Safety* **11**, 378–398 (2012).

45. Lee, S. Y., Baek, N. & Nam, T. G. Natural, semisynthetic and synthetic tyrosinase inhibitors. *Journal of Enzyme Inhibition and Medicinal Chemistry* **31**, 1–13 (2016).

46. Mendes, E., Perry, M. D. J. & Francisco, A. P. Design and discovery of mushroom tyrosinase inhibitors and their therapeutic applications. *Expert Opinion on Drug Discovery* **9**, 533–554 (2014).

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

48. Kohavi, R. & John, G. H. Wrappers for feature subset selection. *Artificial Intelligence* **97**, (1997).

49. Baldi, P., Brunak, S., Chauvin, Y., Andersen, C. A. F. & Nielsen, H. Assessing the accuracy of prediction algorithms for classification: An overview. *Bioinformatics* vol. 16 (2000).

50. Sokolova, M. & Lapalme, G. A systematic analysis of performance measures for classification tasks. *Information Processing and Management* **45**, 427–437 (2009).

51. Valsecchi, C., Grisoni, F., Consonni, V. & Ballabio, D. Consensus versus Individual QSARs in Classification: Comparison on a Large-Scale Case Study. *Journal of Chemical Information and Modeling* **60**, 1215–1223 (2020).

52. Landrum, G. RDKit: Open-source Cheminformatics. *Http://Www.Rdkit.Org/* vol. 3 (2006).

53. Anil, K. T. J. W. Autodock vina: improving the speed and accuracy of docking. *Journal of Computational Chemistry* **31**, (2019).

54. Valdés-Martiní, J. R. *et al.* QuBiLS-MAS, open source multi-platform software for atom- and bond-based topological (2D) and chiral (2.5D) algebraic molecular descriptors computations. *Journal of Cheminformatics* **9**, (2017).

55. May, O. E., Moyer, A. J., Wells, P. A. & Herrick, H. T. The production of kojic acid by aspergillus flavus. *Journal of the American Chemical Society* **53**, (1931).

56. Cui, T., Nakamura, K., Ma, L., Li, J. Z. & Kayahara, H. Analyses of arbutin and chlorogenic acid, the major phenolic constituents in Oriental pear. *Journal of Agricultural and Food Chemistry* **53**, (2005).

57. Deans, B. J., Kilah, N. L., Jordan, G. J., Bissember, A. C. & Smith, J. A. Arbutin Derivatives Isolated from Ancient Proteaceae: Potential Phytochemical Markers Present in Bellendena, Cenarrhenes, and Persoonia Genera. *Journal of Natural Products* **81**, (2018).

58. Albertino, A. *et al.* Natural origin of ascorbic acid: Validation by 13C NMR and IRMS. *Food Chemistry* **112**, (2009).

59. Soedarjo, M. & Borthakur, D. Mimosine produced by the tree-legume leucaena provides growth advantages to some rhizobium strains that utilize it as a source of carbon and nitrogen. in *Plant and Soil* vol. 186 (1996).

60. Gütlein, M., Karwath, A. & Kramer, S. CheS-Mapper - Chemical space mapping and visualization in 3D. *Journal of Cheminformatics* **4**, (2012).

61. Adasme, M. F. *et al.* PLIP 2021: expanding the scope of the protein–ligand interaction profiler to DNA and RNA. *Nucleic Acids Research* (2021) doi:10.1093/nar/gkab294.

62. Atanasov, A. G. *et al.* Natural products in drug discovery: advances and opportunities. *Nature Reviews Drug Discovery* (2021) doi:10.1038/s41573-020-00114-z.

63. Ismaya, W. T. *et al.* Crystal structure of agaricus bisporus mushroom tyrosinase: Identity of the tetramer subunits and interaction with tropolone. *Biochemistry* **50**, (2011).

64. Sendovski, M., Kanteev, M., Ben-Yosef, V. S., Adir, N. & Fishman, A. First structures of an active bacterial tyrosinase reveal copper plasticity. *Journal of Molecular Biology* **405**, (2011).

65. Lai, X., Wichers, H. J., Soler-Lopez, M. & Dijkstra, B. W. Structure of Human Tyrosinase Related Protein 1 Reveals a Binuclear Zinc Active Site Important for Melanogenesis. *Angewandte Chemie - International Edition* **56**, (2017).

66. Ntie-Kang, F. *et al.* AfroDb: a select highly potent and diverse natural product library from African medicinal plants. *PloS one* **8**, 1–15 (2013).

67. Kang, H. *et al.* HIM-herbal ingredients in-vivo metabolism database. *Journal of Cheminformatics* **5**, 1 (2013).

68. Mangal, M., Sagar, P., Singh, H., Raghava, G. P. S. & Agarwal, S. M. NPACT: Naturally occurring plant-based anti-cancer compound-activity-target database. *Nucleic Acids Research* **41**, 1124–1129 (2013).

69. Gražulis, S. *et al.* Crystallography Open Database (COD): An open-access collection of crystal structures and platform for world-wide collaboration. *Nucleic Acids Research* **40**, 420–427 (2012).

7. ANNEXES

TABLES

Table 1. List of 13 tyrosinases and Tyrosinase Related Proteins (TRP) of the panel proposed for docking.

PDB ID	Description	DOI	Year	Resolution	Ligand	Organism	Enzyme form	Metal ions
2y9x	Tyrosinase	10.2210/pdb2Y9X/pdb	2011	2.78	Tropolone	*Agaricus bisporus*	deoxy	Cu
3kan	TRP 2	10.2210/pdb3KAN/pdb	2011	1.13	4IPP	*Homo sapiens*		
3nq1	Tyrosinase	10.2210/pdb3NQ1/pdb	2010	2.3	Kojic Acid	*Bacillus megaterium*	met	Cu, Zn
3w6w	Tyrosinase	10.2210/pdb3W6W/pdb	2013	1.39		*Aspergillus oryzae*	holo-pro	Cu
5ce9	Tyrosinase	10.2210/pdb5CE9/pdb	2015	1.8		*Juglans regia*	met	Cu
5i38	Tyrosinase	10.2210/pdb5I38/pdb	2016	2.6	Kojic acid	*Bacillus megaterium*	met	Cu
5m8m	TRP 1	10.2210/pdb5M8M/pdb	2017	2.65	Kojic Acid	*Homo sapiens*		Zn
5m8o	TRP 1	10.2210/pdb5M8O/pdb	2017	2.5	Tropolone	*Homo sapiens*		Zn
5m8q	TRP 1	10.2210/pdb5M8Q/pdb	2017	2.85	Kojic Acid	*Homo sapiens*		Zn
5z0d	Tyrosinase	10.2210/pdb5Z0D/pdb	2018	1.16		*Streptomyces castaneoglobisporus*	deoxy	Cu
5zrd	Tyrosinase	10.2210/pdb5ZRD/pdb	2018	2.3		*Burkholderia thailandensis*		Cu
6els	Tyrosinase	10.2210/pdb6ELS/pdb	2019	1.35		*Malus domestica*		Cu
6qxd	Tyrosinase	10.2210/pdb6QXD/pdb	2019	2.32	JKB	*Bacillus megaterium*		Cu
PM0079416	Model of human tyrosinase (*)	N/A	2014			*Homo sapiens*	met	

Method: X-ray diffraction excepting (*)

Table 2. Summary of the data used for building and evaluation of the tyrosinase panel and the predictive models

DATA	Molecules	Comments	Activity[a]	Model type[b]	References
D1	36	The data was divided into 3 categories for classifier models to resemble the original paper.	IC50	3D QSAR pharmacophore models	[7]
D2	26		IC50	3D QSAR Pharmacophore-Based Virtual Screening	[41]
D3	20		%Ih 20ug	3D QSAR	[9]
D4	9		IC50	Molecular docking	[22]
D5	44		Ki	Ensemble-Based Virtual Screening	[42]
D6	48		IC50	Pharmacophore models and virtual screening	[18]
D7	56		IC50	3D-QSARs	.[19]
D8	43		IC50	QSARs	[20]
D9	15		%Ih 50 uM	Molecular Docking	[43]
D10	9		IC50	Molecular Docking s	[24]
D11	17		IC50	Pharmacophore model	[17]
D12	33		IC50	3D-QSARs	[21]
D13	93	Review paper	IC50		[28]
D14	1422		IC50	QSARs	[26,27]
D15	1813		IC50		This study

(a)All the cases models of classification and regression were developed except D5 AND D10

(b) Modesl developed in the original studies

Table 3 : Summary of the Screening Data with the virtual hits and cluster analysis results

Screening Data	Name Reference	Molecules	3D Structures No Generated	Molecules Do Not Pass The Time Out 10 Min	Docked	Virtual Hits	Virtual Hits After First Cluster Analysis	Reference
S1	Killer and other collections	800	41	48	711	75(16)*	18(2)	http://www.msdiscovery.com/natprod.html
S2	African Drug Base	885		82	803	56(8)	19(1)	[66]
S3	HIM-herbal	663		126	537	69(11)	15(0)	[67]
S4	NPACT	1423		222	1201	89(16)	16(1)	[68,69]
S5	Nubbe	588		94	494	81(14)	24(1)	https://nubbe.iq.unesp.br/portal/nubbe-search.html
S6	Specs	1496		158	1338	117(17)	30(2)	https://www.specs.net/
S7	INDOFINE	144		23	121	11(3)	3(1)	https://indofinechemical.com/
S8	Selleck Natural Products	144		14	130	12(3)	4(0)	https://www.selleckchem.com/screening/natural-product-library.html
Total		6143	41	767	5335	512(88)	129(8)	
Average		767.875	5.125	95.875	666,875	64(11)	16.125(1)	

*****The number in parentesis are referred to the molecules that were selected as active for the second rule

Table 4: Summary of the results of the models with test of 10 cross-validation

		Regression Models					Classification Models				
Data	**Model variant**	**R^2**	**MAE**	**Machine Learning Technique**	**Docking Parameters**	**Tyrosinase proteins**	**Q(%)**	**MCC**	**Machine Learning Technique**	**Docking descriptors**	**Tyrosinase proteins**
D1	A	0.8898	28.2814	IBK	3	2	91.6667	0.868	IBK	3	3
	B	0.8546	40.7042	SMOreg	6	4	94.4444	0.914	Stacking Classifiers: Logistic, IBK, 1R Metaclassifier: LDA	3	3
D2	A	0.6798	6.6499	SMOreg	6	4	92.3077	0.846	Stacking Classifiers: NaiveBayes, FLDA, Logistic Metaclassifier: NaiveBayes	6	5
	B	0.5513	8.2089	Linear Regression	7	5	88.4615	0.772	Random Forest	2	2
D3	A	0.7874	2.1567	SMOreg	4	3	100	1	FLDA	6	5
	B	0.8059	2.1268	Additive Regression with Gaussian Processes	4	3	100	1	Logistic	6	4
D4	A	0.9598	3.855	IBK	4	4	100	1	IBK	2	2
	B	0.9471	3.4726	SMOreg	2	2	100	1	FLDA	2	2
D5	A	0.7501	2.3519	SMOreg	8	7					
	B	0.7822	2.2829	Additive Regression with Gaussian Processes	8	7					
D6	A	0.8508	37.4137	Random Forest	2	2	87.5	0.507	IBK	3	3
	B	0.8515	34.1298	Additive Regression with	2	2	89.5833	0.579	AdaBoost M1 with Random Forest	3	3

				Random Forest							
D7	A	0.7944	4543.5528	IBK	2	2	100	1	Logistic	6	6
	B	0.8034	4945.2652	SMOreg	9	6	98.2143	0.960	FLDA	8	7
D8	A	0.7423	15.653	SMOreg	6	4	76.7442	0.560	IBK	3	3
	B	0.7347	16.0873	SMOreg	8	6	81.3953	0.640	Stacking Classifier	3	3
D9	A	0.8102	8.7793	SMOreg	5	4	93.3333	0.853	Random Forest	1	1
	B	0.8093	8.1805	Additive Regression with Gaussian Processes	5	4	100	1	Adaboost M1 with LDA	5	5
D10	A	0.7131	49.2231	SMOreg	3	3					
	B	0.7383	45.7662	Bagging with SMOreg	3	3					
D11	A	0.9266	12.231	SMOreg	4	3	100	1	FLDA	5	4
	B	0.9571	9.5399	SMOreg	3	2	94.1176	0.887	IBK	5	5
D12	A	0.8745	1.6529	Staking Classifiers Gaussian Processes, SMOreg, IBK Metaclassifier: IBK	7	5	100	1	FLDA	5	3
	B	0.8734	1.5307	Staking ClassifiersLinear Regression, SMOreg, MP5 Metaclassifier: SMOreg	7	5	96.9697	0.696	Logistic	1	1
D13	A	0.8052	30.1996	SMOreg	7	6	80.6452	0.610	Random Forest	5	4

	B	0.8037	29.7797	Staking Classifiers Gaussian Processes, IBK, MP5 Metaclassifier: MP5	9	7	78.4946	0.568	SMO	13	6
D14	A						79.4655	0.589	Stacking Classifiers: FLDA, IBK, Random Forest Metaclassifier: Logistic	21	12
	B						79.2546	0.586	Stacking Classifiers: SMO, IBK, Random Forest Metaclassifier: Logistic	17	11

Table 5: Statistics of the best base and consensus models of the validation with D15 taking in consideration accuracy, kappa statistic, and F1 Active Score for the 3 breakpoints

Models/IC_{50}	Accuraccy	Correctly Classified Instances	Incorrectly Classified Instances	Kappa statistic	Average Recall	Average Precision	Recall -> Inactive	Recall -> Active	Precision -> Inactive	Precision -> Active	F1 precision
Breakpoint of 1uM											
D10RB	0.846	1533	280	0	0.5	0.423	1	0	0.846	0	#
D3RA	0.837	1517	296	0.02	0.506	0.549	0.984	0.029	0.847	0.25	0.051971
D7CB	0.71	1287	526	0.082	0.55	0.536	0.781	0.318	0.862	0.21	0.252955
D13CB	0.547	992	821	0.053	0.548	0.525	0.547	0.55	0.869	0.181	0.272367
D8CA	0.413	748	1065	0.032	0.537	0.522	0.357	0.718	0.874	0.169	0.273601
D15_P01A	0.782	1417	396	0.12	0.557	0.564	0.882	0.232	0.863	0.264	0.404
D15_P01B	0.697	1264	549	0.114	0.574	0.549	0.752	0.396	0.872	0.226	0.359
Breakpoint of 10uM											
D10RB	0.629	1141	672	0	0.5	0.315	1	0	0.629	0	#
D3RA	0.628	1139	674	0.009	0.504	0.55	0.985	0.022	0.631	0.469	0.042029
D7CB	0.614	1113	700	0.104	0.548	0.563	0.802	0.295	0.659	0.467	0.361588
D7CA	0.58	1052	761	0.101	0.55	0.55	0.666	0.435	0.667	0.434	0.434499
D12RA	0.371	672	1141	0	0.5	0.185	0	1	0	0.371	0.541211
D12CB	0.397	719	1094	0.017	0.511	0.546	0.068	0.954	0.716	0.376	0.539405
D15_P10A	0.65	1179	634	0.141	0.562	0.619	0.904	0.219	0.663	0.574	0.615
D15_P10A	0.601	1090	723	0.162	0.583	0.58	0.655	0.51	0.694	0.465	0.557
Breakpoint of 30uM											
D8CA	0.547	991	822	0.097	0.549	0.554	0.393	0.704	0.577	0.531	0.553
D7CB	0.544	986	827	0.081	0.54	0.556	0.806	0.275	0.533	0.58	0.556
D12RA	0.494	895	918	0	0.5	0.247	0	1	0	0.494	0.661312
D12CB	0.514	932	881	0.039	0.52	0.587	0.08	0.96	0.67	0.504	0.660984
D15_P30A	0.549	995	818	0.104	0.553	0.582	0.253	0.853	0.637	0.527	0.577
D15_P30B	0.591	1072	741	0.18	0.59	0.594	0.691	0.489	0.581	0.607	0.594

FIGURES

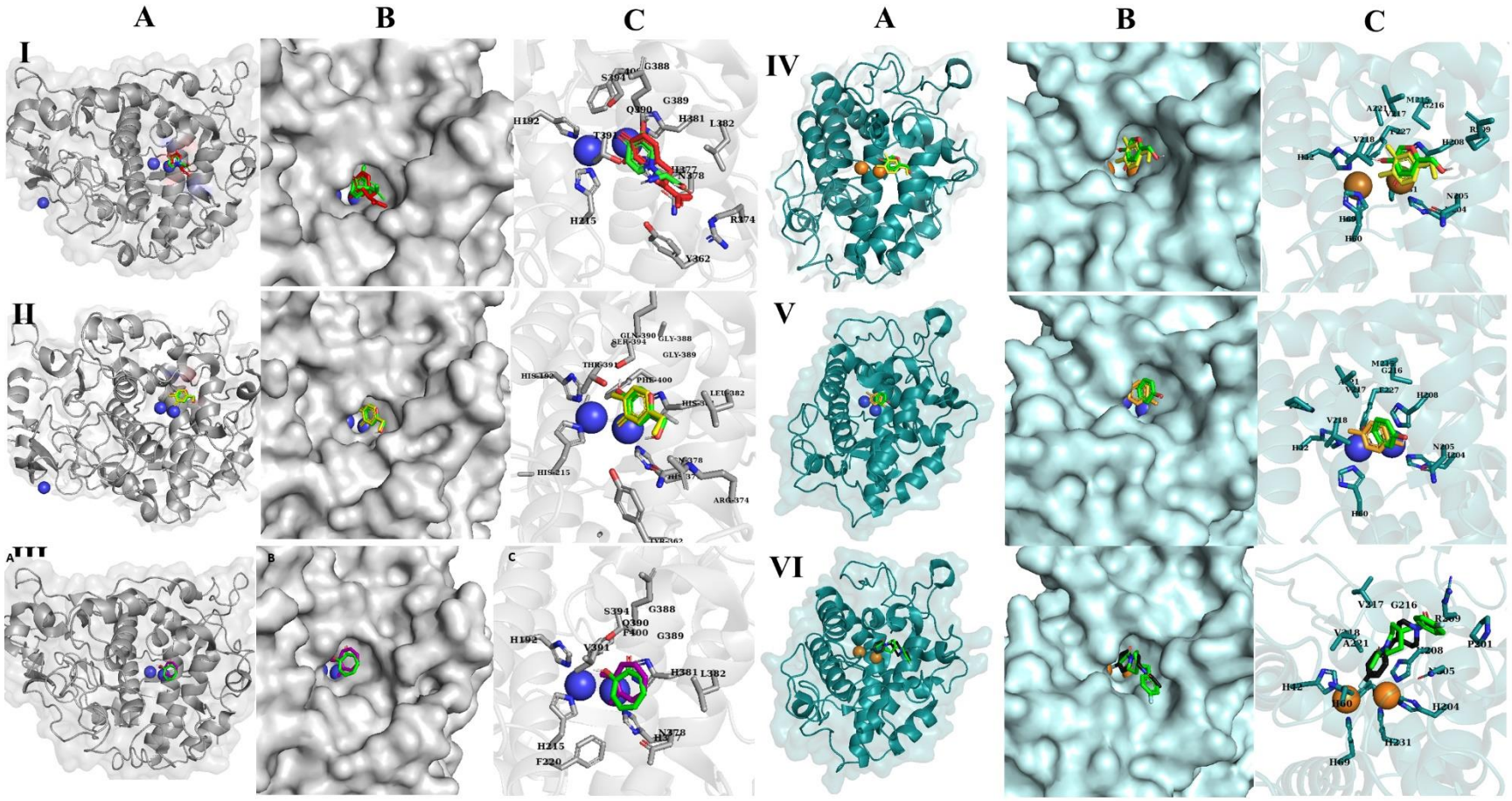


Figure 1. Representation of the re-docking runs of TRP1 of *Homo sapiens* (grey) and Tyrosinase of *Bacillus megaterium* (turquoise). Crystalized conformation of mimosine, kojic acid (KA), tropolone, hydroquinone, SVF is shown in red, yellow, purple, orange, and black, respectively. The best-docked pose is depicted in green. The zinc and copper ions are presented as blue and orange spheres, respectively. I) 5M8N chain A complex-mimosine, run 1, pose 1, dock affinity -6.1, and RMSD strict 1.21 II) 5M8M chain A complex-KA, run 3, pose 7, dock affinity -5.0, and RMSD strict 0.25. III) 5M8T chain A complex-tropolone, run 2, pose 8, dock affinity -5.4, and RMSD strict 0.4. IV) 5I38 chain A complex-KA, run 3, pose 5, dock affinity -5, and RMSD strict 0.76. V) 5I3B chain A complex-hydroquinone, run 2, pose 4, dock affinity -5.1, and RMSD strict 1.29. VI) 5OAE chain A complex-SVF, run 3, pose 9, dock affinity -6.4, and RMSD strict 2.07. A) Global cartoon representation B) Surface representation C) Residues in the interacting at 5 Å

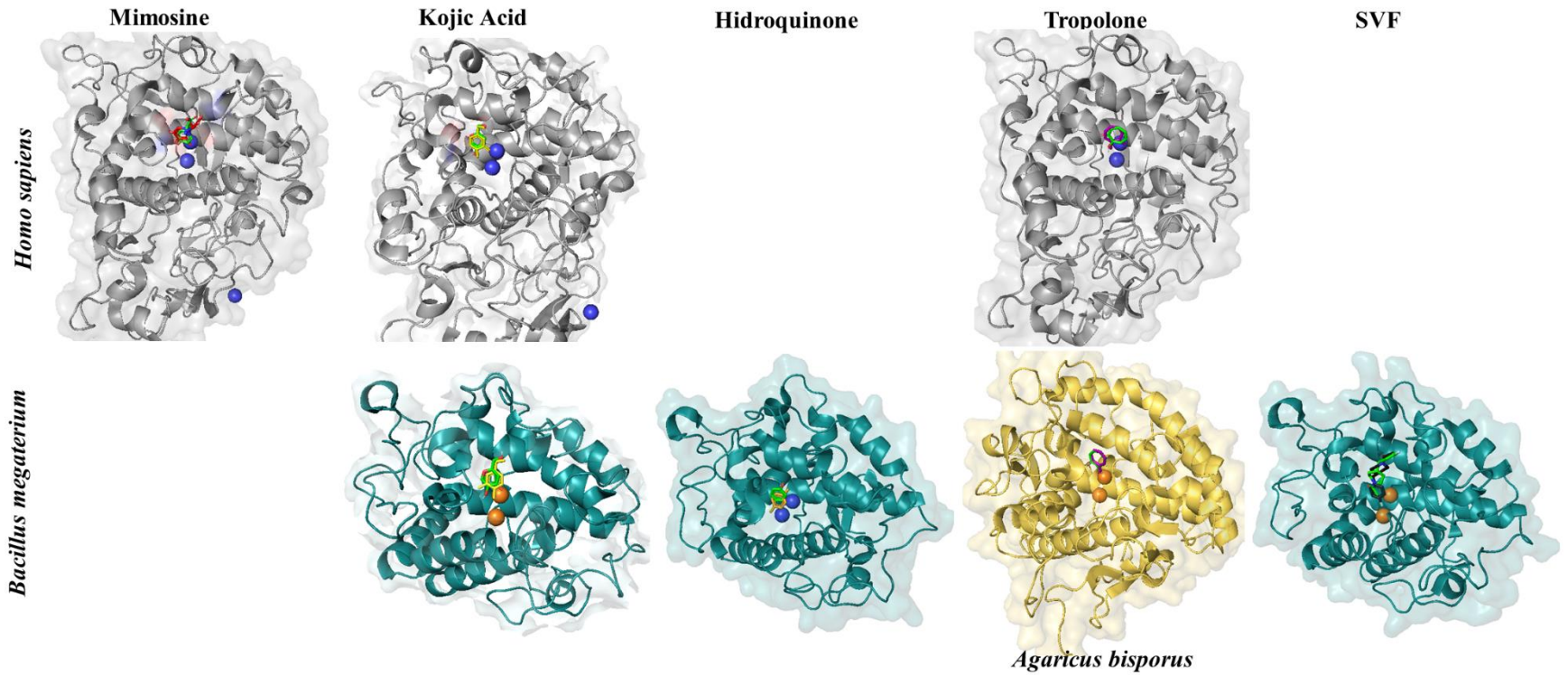


Figure 2. Representation of Tyrosinase and TRP with several ligands. TRPs cartoons from *H. sapiens* are in grey (PDB ID: 5M8T, 5M8M, 5M8M), from *B. megaterium* are in cyan (PDB ID: 5OAE, 5I3B, 5I38), and from *A. bisporus* is in yellow (PDB ID: 2Y9X). The zinc and copper ions are presented as blue and orange spheres, respectively. The co-crystalized positions are in green.

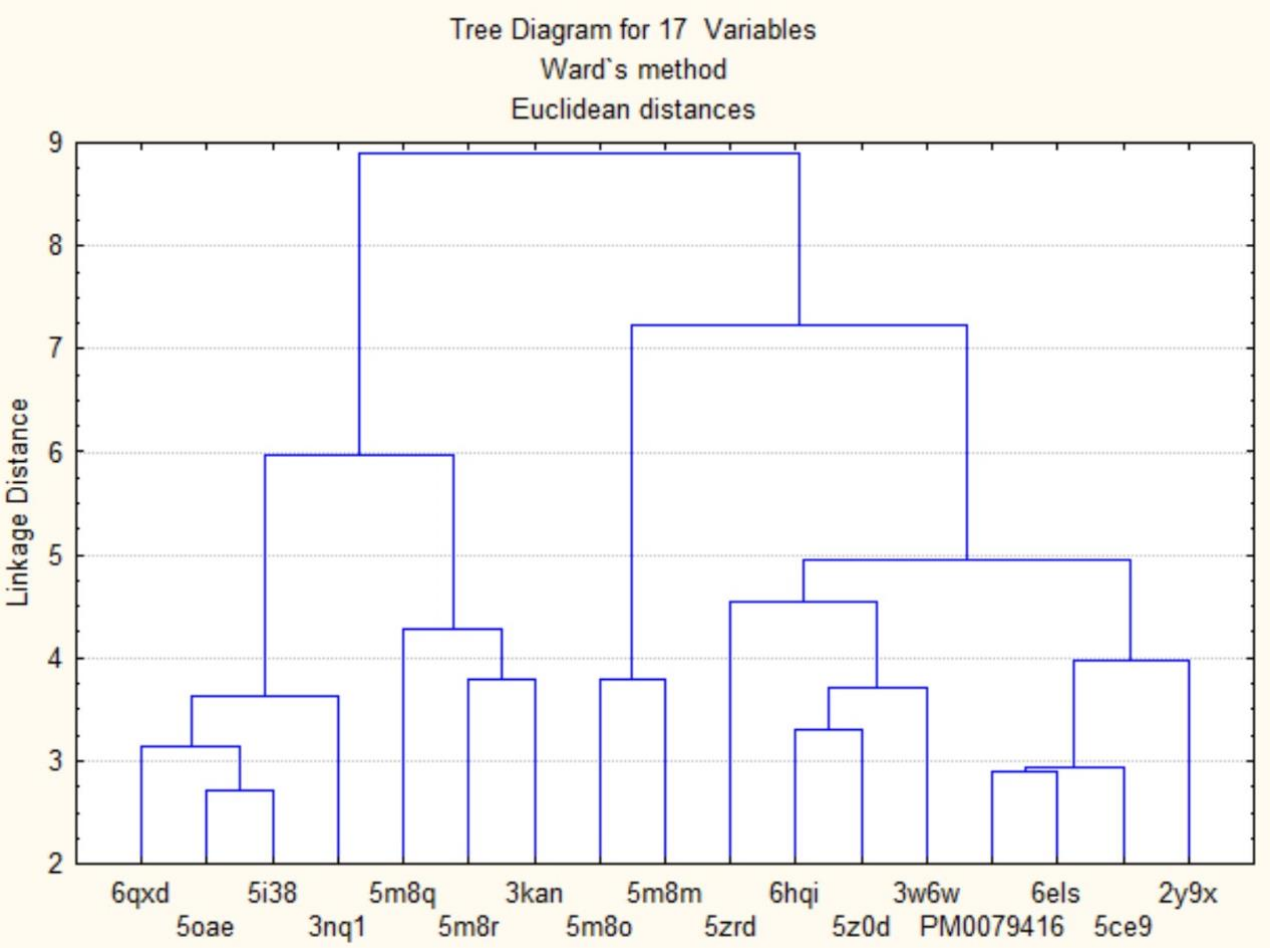


Figure 3. Dendrogram illustrating the results of the CA of the 17 tyrosinases and TRPs. From this CA, 13 PDBs were selected as representative and diverse: 2Y9X, 3KAN, 3NQ1, 3W6W, 5CE9, 5I38, 5M8M, 5M8Q, 5Z0D, 5ZRD, 6ELS, 6QXD, and PM0079416.

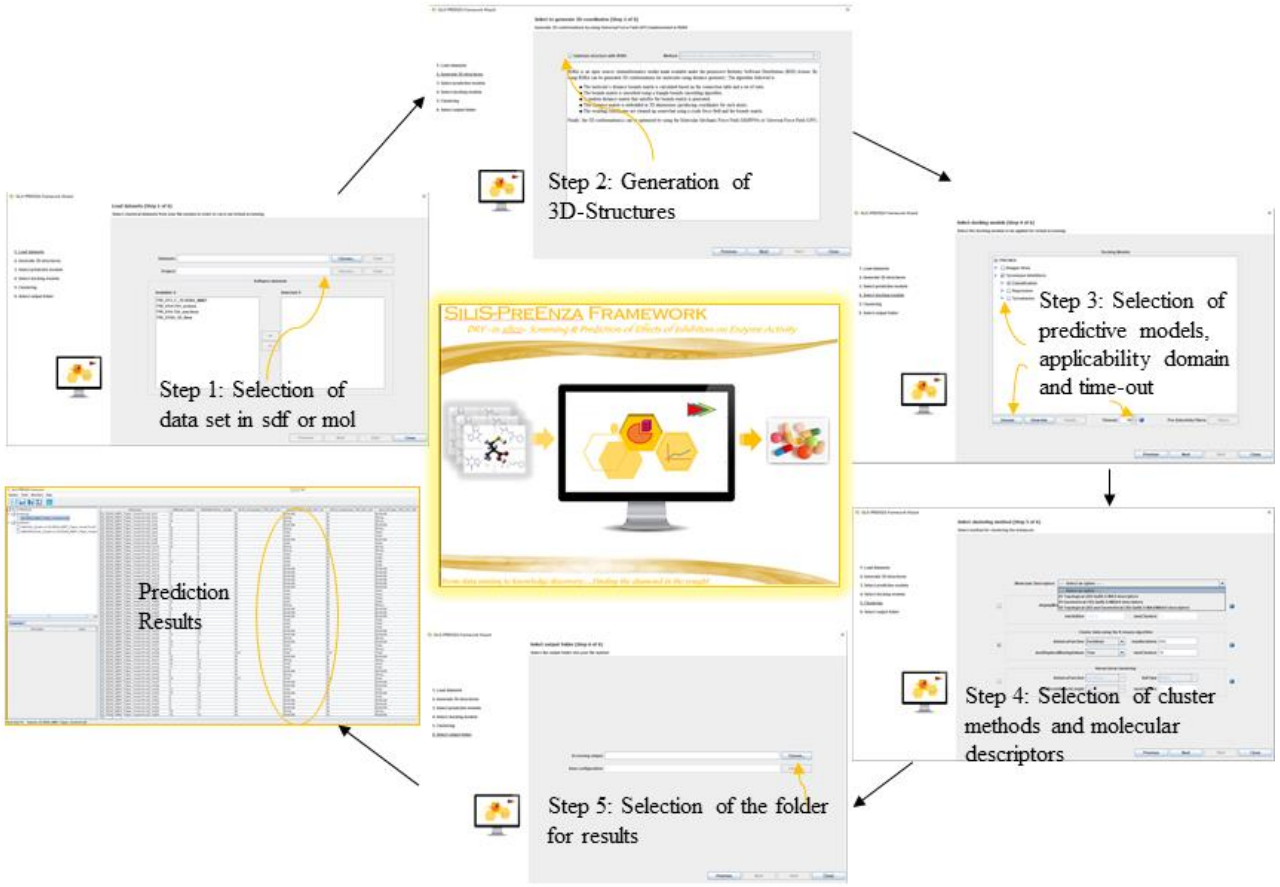


Figure 4. Screenshots of the SiliS-PREENZA software: (step 1) interface to select compounds in SDF or MOL files; (step 2) generation of 3D structures in the case the data is in 2D; (step3) interface to select and show the information of the predictive model(s) to be used of classification or regression, as well as to select the proteins for docking (s). There could also select to compute the applicability domain(s) and the time-out function; (step 4) Selection of the clustering method(s) and the descriptors used for it; and (step 5) interface for the processing of the results obtained.

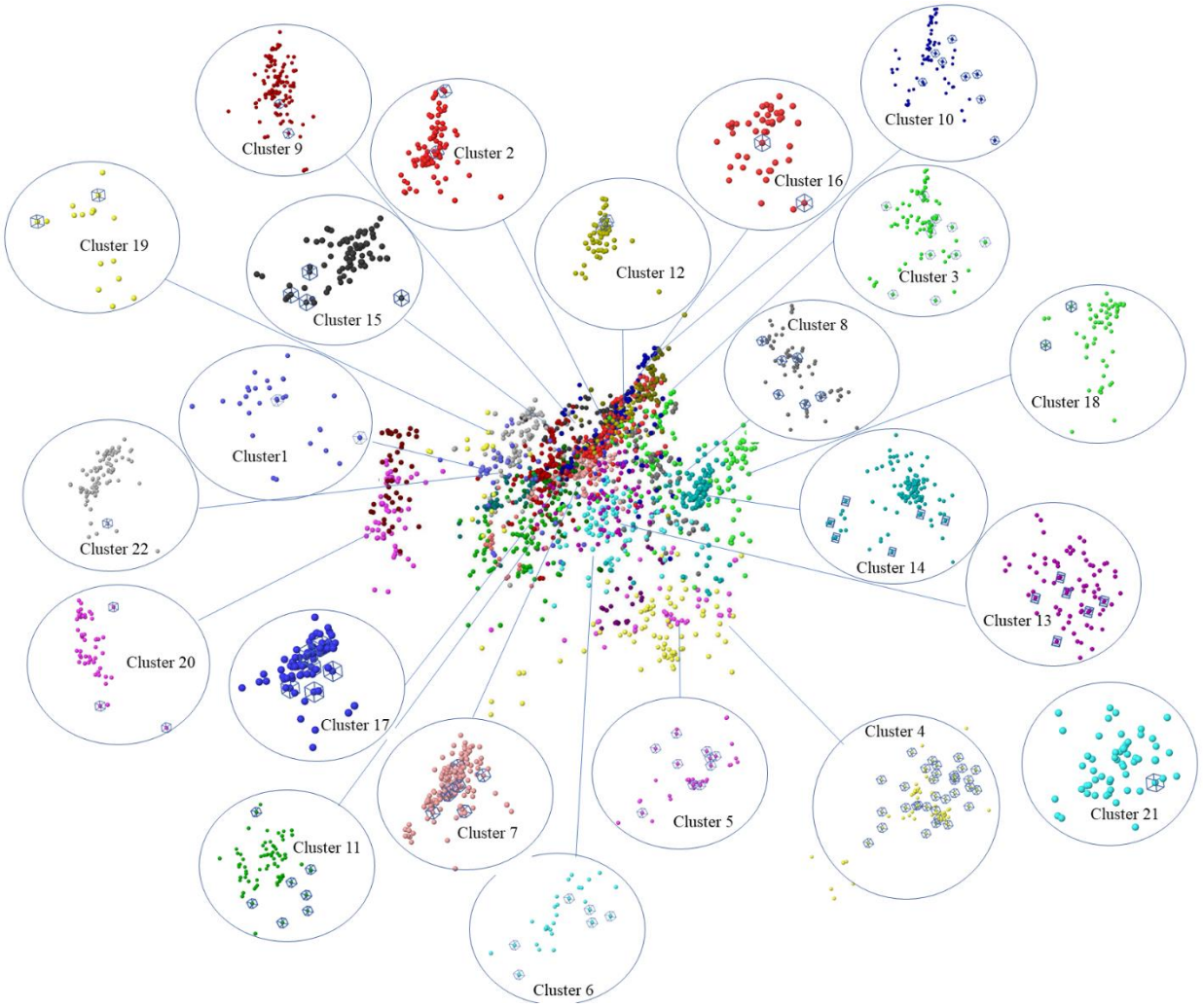


Figure 5. Ches-S clustering in 30 groups of the 131 virtual hits selected and the 2514 reference molecules of TIs. The Virtual hits were compared and grouped into 22 clusters. The colors indicate compounds grouped by structural difference. The compounds marked inside each cluster are the virtual hits.

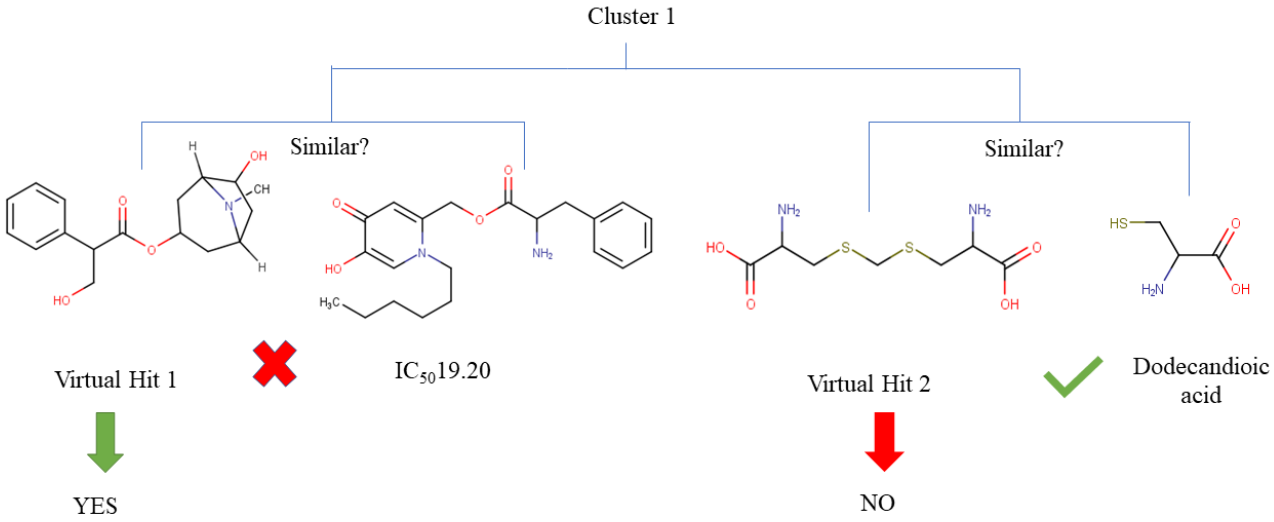


Figure 6. Pseudo-dendrogram (decision-making process) of virtual hits and their closest neighbor of Cluster 1.

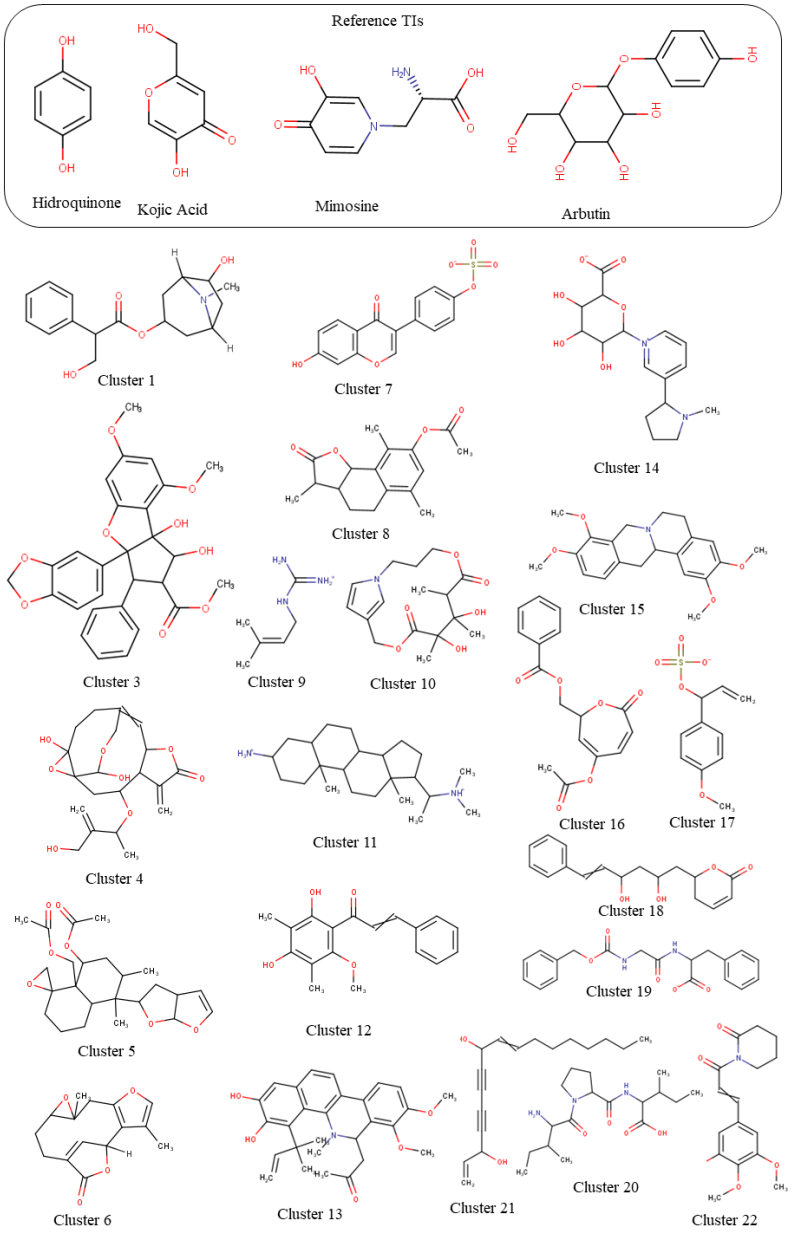


Figure 7. Schematic representation of the TIs more representative and the main virtual hits of each cluste

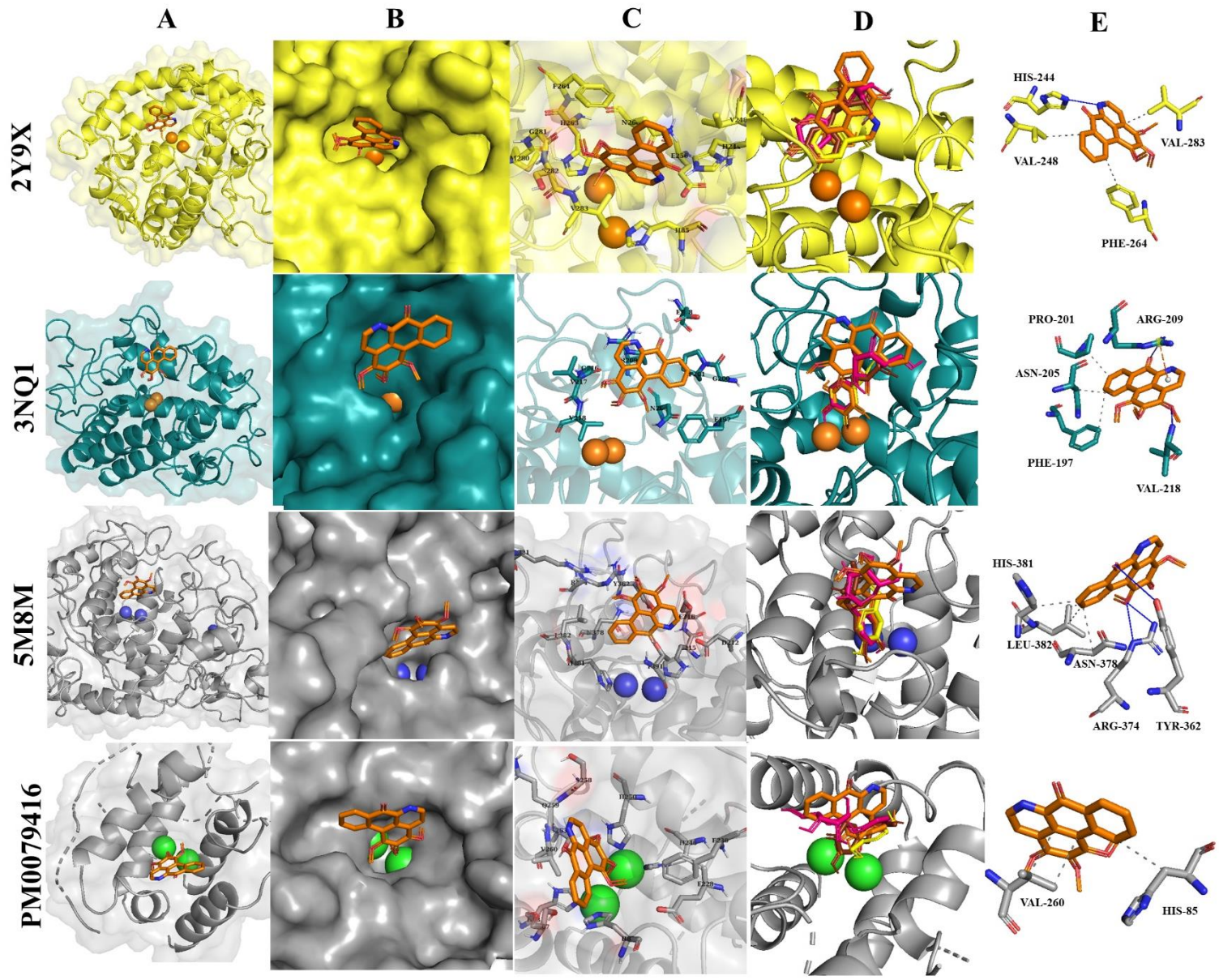

A
B
C
D
E
2Y9X
3NQ1
5M8M
PM0079416
HIS-244
VAL-283
VAL-248
PHE-264
PRO-201
ARG-209
ASN-205
PHE-197
VAL-218
HIS-381
LEU-382
ASN-378
ARG-374
TYR-362
VAL-260
HIS-85

Figure 8: Representation of the virtual lead Liridine ID: ZINC0033812 docked against tyrosinases from *Agaricus bisporus* (PDB:2Y9X), *Bacillus megaterium* (PDB:3NQ1), *Homo sapiens* (PDB:5M8M and PM0079416) the last ones are TRP1 and a model of human tyrosinase respectively. Docked conformations of liridine, Arbutin, hydroquinone, kojic acid is shown in bright orange, hotpink, orange, yellow and brown respectively. A) Global cartoon representation B) Surface representation C) Residues in the interacting at 4 Å D) virtual lead with for 4 known tyrosinase inhibitors E) Representation of the molecular forces interacting with the lead compound