UNIVERSIDAD DE INVESTIGACIÓN DE TECNOLOGÍA EXPERIMENTAL YACHAY

Escuela de Ciencias Biológicas e Ingeniería

TÍTULO: Prediction Models for the Discovery of Insect Repellents that Interfere with Olfaction

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

Autora:

Mary Estefanía Pulgar Sánchez

Tutor:

Ph.D. Markus P. Tellkamp

Co-tutor:

Ph.D. Yovani Marrero – Ponce

Urcuquí, Junio 2021.

Urcuquí, 11 de junio de 2021

SECRETARÍA GENERAL
(Vicerrectorado Académico/Cancillería)
ESCUELA DE CIENCIAS BIOLÓGICAS E INGENIERÍA
CARRERA DE BIOLOGÍA
ACTA DE DEFENSA No. UITEY-BIO-2021-00012-AD

A los 11 días del mes de junio de 2021, a las 15:30 horas, de manera virtual mediante videoconferencia, y ante el Tribunal Calificador, integrado por los docentes:

Presidente Tribunal de Defensa	Dr. SANTIAGO VISPO, NELSON FRANCISCO , Ph.D.
Miembro No Tutor	Dra. SPENCER VALERO, LILIAN MARITZA , Ph.D.
Tutor	Dr. TELLKAMP TIETZ, MARKUS PATRICIO , Ph.D.

El(la) señor(ita) estudiante **PULGAR SANCHEZ, MARY ESTEFANIA**, con cédula de identidad No. **0604449157**, 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: **Prediction models for the discovery of insect repellents that interfere with olfaction** , previa a la obtención del título de **BIÓLOGO/A**.

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

Tutor	Dr. TELLKAMP TIETZ, MARKUS PATRICIO , 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
Tutor	Dr. TELLKAMP TIETZ, MARKUS PATRICIO , Ph.D.	9,8
Miembro Tribunal De Defensa	Dra. SPENCER VALERO, LILIAN MARITZA , Ph.D.	10,0
Presidente Tribunal De Defensa	Dr. SANTIAGO VISPO, NELSON FRANCISCO , Ph.D.	10,0

Lo que da un promedio de: **9.9 (Nueve punto Nueve)**, 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.*

MARY ESTEFANIA PULGAR SANCHEZ
Firmado digitalmente por MARY ESTEFANIA PULGAR SANCHEZ
Fecha: 2021.06.21 12:07:45 -05'00'

PULGAR SANCHEZ, MARY ESTEFANIA
Estudiante

NELSON FRANCISCO SANTIAGO VISPO
Firmado digitalmente por NELSON FRANCISCO SANTIAGO VISPO
Fecha: 2021.06.11 16:43:12 -05'00'

Dr. SANTIAGO VISPO, NELSON FRANCISCO , Ph.D.
Presidente Tribunal de Defensa

Firmado electrónicamente por:
MARKUS PATRICIO TELLKAMP TIETZ

Dr. TELLKAMP TIETZ, MARKUS PATRICIO , Ph.D.
Tutor

LILIAN MARITZA SPENCER VALERO

Firmado digitalmente por LILIAN MARITZA SPENCER VALERO
Fecha: 2021.06.14 15:24:50 -05'00'

Dra. SPENCER VALERO, LILIAN MARITZA , Ph.D.
Miembro No Tutor

KARLA ESTEFANIA ALARCON FELIX

Firmado digitalmente por KARLA ESTEFANIA ALARCON FELIX
Fecha: 2021.06.11 16:52:27 -05'00'

ALARCON FELIX, KARLA ESTEFANIA
Secretario Ad-hoc

AUTORÍA

Yo, **Mary Estefanía Pulgar Sánchez**, con cédula de identidad 060444915-7, 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, Junio, 2021.

MARY ESTEFANIA PULGAR SANCHEZ

Firmado digitalmente por MARY ESTEFANIA PULGAR SANCHEZ
Fecha: 2021.06.22 12:14:36 -05'00'

Mary Estefanía Pulgar Sánchez

CI: 060444915-7

AUTORIZACIÓN DE PUBLICACIÓN

Yo, **Mary Estefanía Pulgar Sánchez**, con cédula de identidad 060444915-7, 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, Junio, 2021.

MARY ESTEFANIA PULGAR SANCHEZ

Firmado digitalmente por MARY ESTEFANIA PULGAR SANCHEZ
Fecha: 2021.06.22 12:15:04 -05'00'

Mary Estefanía Pulgar Sánchez

CI: 060444915-7

Dedicatoria

Para mis padres, Rosario y Rómulo, que son los mejores y siempre han sido mi apoyo incondicional en cada paso que doy. A su eterno amor, paciencia, entrega y esfuerzo que hacen a diario por la familia. Son lo más bonito de mi vida.

Para mis hermanas Jessy y Karito por siempre cuidar de mí; y por supuesto para mis amados sobrinos, Francisco y Analía, por alegrar mis días con su ternura.

Mary Estefanía Pulgar Sánchez

Acknowledgments

I want to express my gratitude to my co-advisor Yovani Marrero-Ponce, whose support and scientific guidance helped develop this research. Special thanks to Markus Tellkamp for the support in this thesis and all the projects we worked on. I also thank the programming team that helped me in the implementation of the software.

I want to extend a fraternal thank you to my Yachay Tech teachers for being so amazing, sharing their knowledge, and being our inspiration. I also thank my best friend Rosita for always being with me in my happy and challenging times. Besides, I acknowledge my bio friends Leandro and Camila, for the teachings, adventures, and support. And finally, thanks to all my friends from Yachay Tech, especially appreciation to Danii, Amanda, Migue, and my entire team who became my family and helped turn my moments of stress and sadness into laughter.

This accomplishment would not have been possible without the support of all of them.

Mary Estefanía Pulgar Sánchez

Resumen

Los insectos vectores de enfermedades dependen quimiorreceptores para localizar a los huéspedes, encontrar parejas y elegir dónde poner sus huevos. Actualmente, el método más eficaz para prevenir y controlar el brote de enfermedades transmitidas por insectos ha sido el uso de repelentes de insectos (RIs). Sin embargo, no cumplen con las condiciones necesarias como para brindar protección ante un amplio espectro de mosquitos; muchos de ellos tienen olor o sensaciones en la piel desagradable, incluso algunos de ellos son cancerígenos; es decir, los RI actuales tienen importantes inconvenientes. Por lo tanto, es evidente la necesidad de nuevos RI de protección de amplio espectro, más efectivos, seguros y duraderos que los RI convencionales. Aquí, los clasificadores para la predicción de RI se desarrollarán mediante el uso de descriptores moleculares QuBiLS Suite 0-3D y técnicas de aprendizaje automático superficial. Los mejores modelos individuales se usaron para obtener modelos conjuntos con parámetros estadísticos adecuados para la serie de aprendizaje. En el presente trabajo pretendemos introducir, por primera vez, la capacidad de QSAR- (Relaciones cuantitativas estructura-actividad) y modelos basados en estructura para describir la interacción de los RI con la respuesta olfativa de la sensilla del mosquito *Culex quinquefasciatus*, así como con actividades repelentes mediante el uso de cuatro conjuntos de datos que tomaron en consideración los dos andamios de RI más relevantes: carboxamidas y compuestos derivados de plantas con efecto repelente sobre *A. aegypti* (y también *A. gambiae*) y las dos especies más comunes de cucarachas (*Blattella germanica* y *Periplaneta americana*). Se realiza un software no comercial y multiplataforma, denominado “SiliS-PAPACS”, para la predicción de RI, que está disponible gratuitamente en http://tomocomd.com/apps. Este software se utilizará para el cribado de conjuntos de datos que contienen diversos quimiotipos como componentes de aceites esenciales, productos químicos y medicamentos aprobados por la FDA. El propósito es evaluar la utilidad de los modelos desarrollados en el etiquetado IR de sustancias orgánicas y mostrar la capacidad del sistema para identificar nuevos andamios químicos. Aquí presentamos 23 compuestos nuevos encontrados a través de cribado virtual que pueden tener potencial actividad repelente. Los resultados sugieren que el método propuesto será un buen sistema asistido por computadora que podría aumentar potencialmente la posibilidad de encontrar nuevos agentes para el control de insectos. Es decir, este estudio proporciona información importante que ayudará a los que evalúan y/o diseñan nuevos quimiotipos de RIs.

Palabras Clave: Repelente de Insectos, Proteína de Unión a Odorantes, Software SiliS-PAPACS; índices algebraicos 0-3D basados en átomos y enlaces, aprendizaje automático, QSAR, docking.

Abstract

Disease vector insects rely on chemosensors to locate hosts, find mates and choose where to lay their eggs. Currently, the most efficient method of preventing and controlling the outbreak of insect-borne diseases has been the use of insect repellents (IRs). However, they do not meet the necessary conditions, such as protecting a broad spectrum of mosquitoes; many of them have unpleasant odors or produce unpleasant sensations on the skin, some of them are even carcinogens. In other words, current IRs have significant drawbacks. Therefore, the need for new, more effective, safer, and longer-lasting broad-spectrum IRs than conventional IRs is evident. Here, classifiers for predicting IRs will be developed by using QuBiLS Suite 0-3D molecular descriptors and shallow machine learning techniques. The best individual models were used to obtain ensemble models with suitable statistical parameters for the learning series. In the present work, we intend to introduce, for the first time, the ability of QSAR- (Quantitative Structure-Activity Relationships) and structure-based models to describe the interaction of IRs with the olfactory response of the sensilla of the mosquito *Culex quinquefasciatus* as well as with repellent activities by using four datasets that take into consideration the two most relevant IR scaffolds: carboxamides and plant-derived compounds with repellent effect on *A. aegypti* (and also *A. gambiae*) and the two most common species of cockroach (*Blattella germanica* and *Periplaneta americana*). A non-commercial and cross-platform software named "SiliS-PAPACS" was developed for the IRs-prediction and is freely available at http://tomocomd.com/apps. This software will be used for the screening of datasets containing diverse chemotypes like essential oils constituents, chemicals, and FDA-approved drugs. The purpose is to assess the usefulness of the developed models in the IR-labeling of organic substances and show the system's ability to identify novel IR leads (new IR chemical Scaffold). Here, we report 23 novel compounds found through virtual screening that may have potential repellent activity. The results suggest that the proposed method will be an excellent computer-assisted system that could increase the chance of finding new insect control agents and assist those researchers in screening and/or designing new chemotype IRs.

Keywords: Insect Repellent, Odorant Binding Protein, SiliS-PAPACS Software; Atom- and Bond-based 0-3D Algebraic Indices, Machine Learning, QSAR, Docking.

Contents

List of Figures

List of Tables

LIST OF ABBREVIATIONS

AD: applicability domain

AV: Autodock Vina

BN: Bayes Network algorithm

CV: cross-validation

DEET: *N*, *N*-Diethyl-3-methyl-benzamide

EOC: Essemtial oil constituent

EXP: experiment

FLDA: Fisher's Linear Discriminant function algorithm

GP: Gaussian Processes algorithm

GP: grooved peg

IBk: Instance Based Learner algorithm

IR: insect repellent

KNN: K-nearest neighbors

LDA: Linear Discriminant Analysis algorithm

Log: Logistic regression function algorithm

LR: Linear Regression algorithm

LST: long sharp trichoid

MAE: mean absolute error

MCC: Matthews correlation coefficient

MD: molecular descriptor

MED: minimum effective dose

ML: Machine learning

MLR: Multiple linear regression

NB: Naïve Bayes algorithm

OBP: odorant binding protein

OR: odorant receptor

ORN: odorant receptor neuron

PDB: Protein Data Bank

PR: percentage of repellency

PUK: Pearson Universal Kernel

Q: accuracy

q^2: the R-squared value obtained from applying a QSAR model to the test set instead of the training set

QDA: Quadratic Discriminant Analysis algorithm

QSAR: Quantitative structure-activity relationship

R^2: squared correlation coefficient

RD_{50}: exposure concentration producing a 50% respiratory rate decrease

RF: Random Forest tree algorithm

RMSD: root-mean-square deviation

SBT: short blunt trichoid

SGD: Stochastic Gradient Descent

SiliS-PAPACS: DRY – *in silico* – Screening & Prioritization of Anti-Parasitics and Agro-Chemicals Software

SMO: John Platt's Sequential Minimal Optimization algorithm

SMOreg: Sequential Minimal Optimization Regression algorithm

SST: short sharp trichoid

VOC: volatile organic compound

VS: virtual screening

WEKA: Waikato Environment for Knowledge Analysis

PUBLICATION MANUSCRIPT

TITLE:

Prediction Models for the Discovery of Insect Repellents that Interfere with Olfaction

JOURNAL:

Journal of Computational Chemistry

AUTHORS:

Mary Pulgar-Sánchez

Yovani Marrero-Ponce

ORCID:

https://orcid.org/0000-0001-5966-9065

https://orcid.org/0000-0003-2721-1142

e-mail:

mary.pulgar@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 Computational Chemistry journal format starting next page.

Prediction Models for the Discovery of Insect Repellents that Interfere with Olfaction

ABSTRACT

Disease vector insects rely on chemiosensors to locate hosts, find mates and choose where to lay their eggs. Currently, the most efficient method of preventing and controlling the outbreak of insect-borne diseases has been the use of insect repellents (IRs). However, they do not meet the necessary conditions, such as protecting a broad spectrum of mosquitoes; many of them have unpleasant odors or produce unpleasant sensations on the skin, some of them are even carcinogens. In other words, current IRs have significant drawbacks. Therefore, the need for new, more effective, safer, and longer-lasting broad-spectrum IRs than conventional IRs is evident. Here, classifiers for predicting IRs will be developed by using QuBiLS Suite 0-3D molecular descriptors and shallow machine learning techniques. The best individual models were used to obtain ensemble models with suitable statistical parameters for the learning series. In the present work, we intend to introduce, for the first time, the ability of QSAR- (Quantitative Structure-Activity Relationships) and structure-based models to describe the interaction of IRs with the olfactory response of the sensilla of the mosquito *Culex quinquefasciatus* as well as with repellent activities by using four datasets that take into consideration the two most relevant IR scaffolds: carboxamides and plant-derived compounds with repellent effect on *A. aegypti* (and also *A. gambiae*) and the two most common species of cockroach (*Blattella germanica* and *Periplaneta americana*). A non-commercial and cross-platform software named "SiliS-PAPACS" was developed for the IRs-prediction and is freely available at http://tomocomd.com/apps. This software will be used for the screening of datasets containing diverse chemotypes like essential oils constituents, chemicals, and FDA-approved drugs. The purpose is to assess the usefulness of the

developed models in the IR-labeling of organic substances and show the system's ability to identify novel IR leads (new IR chemical Scaffold). Here, we report 23 novel compounds found through virtual screening that may have potential repellent activity. The results suggest that the proposed method will be an excellent computer-assisted system that could increase the chance of finding new insect control agents and assist those researchers in screening and/or designing new chemotype IRs.

Keywords: Insect Repellent, Odorant Binding Protein, SiliS-PAPACS Software; Atom- and Bond-based 0-3D Algebraic Indices, Machine Learning, QSAR, Docking.

Graphical Abstract

QSAR- & structure- based models

1. INTRODUCTION

1.1. Insects as vectors for diseases

Diseases with epidemic potential are a growing problem worldwide and have been so problematic that, in many cases, health systems have not been able to control it, and it ends up collapsing. Even though insects play essential roles in ecosystems, [1], [2] several species have been responsible for numerous outbreaks of vector-borne diseases that have caused the death of millions of people throughout the world.

So far, the most effective method of controlling this threat to humanity has been through species-specific disruption strategies[2], [3] like the use of pesticides and insect repellents (IRs). The need for IRs has been growing every year. By 1997, it was estimated that 200 million people worldwide use DEET-based IRs per year[4] and in America alone, around 110 million people used IRs[5]. The urgency to develop novel IRs is growing each year as climate change increasingly impacts many insects' distribution.[2], [6]

Chemoception, i.e., sense of smell and taste, can mediate behavioral and physiological responses.[7] Mosquitoes are attracted by odor and estrogens in human sweat, carbon dioxide, moisture, and heat[4]. Female mosquitoes are hematophagous and their main attractant compound is 1-octen-3-ol,[8]–[10] but also by carbon dioxide,[6], [11] acetic acid[12], and lactic acid[13]. Thus the olfactory pathway plays a major role in locating hosts, food, mates, or habitat for laying eggs. This sensory system is an opportunity to interrupt the process of disease transmission.[2], [14] Mosquito bites can be dangerous due to systemic allergic reactions,[4] as well as transmitting parasites and pathogens, resulting in diseases such as malaria, dengue fever, filariasis, leishmaniasis, trypanosomiasis, yellow fever, Chagas disease, Lyme disease, St. Louis encephalitis, Rocky Mountain spotted fever, and many more.[1], [2], [15]–[17]

1.2. Common repellents, mode of action, and related problems

IRs are prophylactic agents that lower the chance of disease transmission by insect vectors[18], [19], by reducing mosquito-human interactions. Several studies have determined the mode of action of IRs. However, it is believed that they primarily disrupt olfactory responses to odor molecules that control mosquito behavior.[14]

The most commonly used repellents are not necessarily the perfect ones. For instance, DEET (*N*, *N*-Diethyl-3-methyl-benzamide), one of the most popular ones, has an unpleasant odor and skin feel and causing a plasticizing effect in some polymers.[4], [20] DEET is also known to be neurotoxic,[21] be subject to declining efficacy with time, cause skin irritation, and be prone to development of resistance.[22], [23] At present, there are two main alternatives to DEET i) IR3535 (3-(*N*-acetyl-*N*-butyl)amino propionic acid ethyl ester), and ii) Picaridin ((RS)-sec-butyl 2-(2-hydroxyethyl)piperidine-1-carboxylate). IR3535 is safer than DEET as evidenced by low acute oral and dermal toxicity and less irritation of mucous membranes.[20] Besides, Picaridin appears to have low acute toxicity and irritating effect.[24]

1.3. The ideal repellent

Therefore, the need for safer IRs is evident. An ideal repellent must be more effective than the actual ones, work on a broad-spectrum of insects, produce longer-lasting protection, and be safer than conventional IRs. That is, being non-irritating to the skin after topical application of safe to use on different kind of clothes, neither staining, bleaching, or weakening them, and being inert and non-reactive with commonly used materials (e.g., plastics, vinyl, spandex, eyeglass frames, pens).[4] Furthermore, it is desired to have a pleasant odor or be odorless, and have a non-greasy feel or appearance and resistance to removal after wiping, washing, or sweating.[20]

1.4. Pheromones and insect attraction

Chemical communication is essential during animal interactions of the same or different species. Insect pheromones are chemical compounds specialized in triggering physiological or behavioral changes in insects of the same species.[25], [26] Pheromones are crucial for insect survival. Different types act in specific ways, for instance, acting as alarm signal; identification of trail, home, royalty, and recruitment pheromones; and sex and oviposition-inhibition pheromones.[27] This type of specific reaction that pheromones produce has been widely used as a pest management strategy.[27] There is also chemical communication between different species through is through allomones that give the signal of repellency and kairomones that act as attractants.[26]

1.5. Proteins involved in insect olfactory system.

One of the study models that stand out in neurosciences is the olfactory system of insects.[2] It has been fundamental for discovering potential targets in vector insects and, therefore, the development of new repellents. Insects' chemical communication processes allow them to detect and differentiate among thousands of odorants in a highly selective manner.[28] In mosquitoes, the olfactory pathways rely on the sensory receptors of antennae and maxillary palps. The olfactory pathways begin upon the reception of volatile organic compounds (VOCs) by specialized hair-like structures in the antennae known as sensilla. Each sensillum has a different shape, being either long sharp trichoid (LST), short sharp trichoid (SST), short blunt trichoid I (SBT-I), short blunt trichoid II (SBT-II), short blunt trichoid-curved (SBT-C), or in the form of a grooved peg (GP).[29]

VOCs are transported to the receptors on the membrane of the sensory neuron where signal transduction takes place. Odorant Binding Proteins (OBPs) are small

water-soluble proteins[30] (10-20 kDa) that are required for the correct performance of the olfactory system. Each OBP may specifically recognize a class of structurally related odorants and also distinguish semiochemicals of different chemical structures.[31]–[33] OBPs direct the transfer of VOCs to the Olfactory Receptors ORs throughout the sensillum aqueous lymph. [34]–[36]. ORs are the most diverse G-protein-coupled receptors (GPCRs) subclass which, in a difference to vertebrates' GPCRs, have its seven transmembrane domains inversely oriented.[37], [38] ORs are located on the membrane of the olfactory receptor neurons (ORNs) and couple an Odorant receptor co-receptor (Orco). Orco protein sequence is highly conserved among insects[2], [39], making it an important target for the development of IRs. Furthermore, Sensory Neuron Membrane Proteins[40] (SNMP), insect-specific ionotropic glutamate receptors (TRPA)[40]–[43] are also part of the olfactory signal transduction pathway at glomeruli of mosquito. Then the information is processed, and a behavioral response is triggered.[7], [44]–[47] OR, Orco, ORN, and OBPs are potential endpoints for controlling insects' behavior to analyze attraction or avoidance circuits.[2]

1.6. QSAR- and docking- based previous theoretical studies of repellency

The drawbacks of current repellents and the increasing interest in designing a natural-based repellent[48] have constantly pushed the research in this field. However, the high cost and long periods of time required for experimental research have hindered progress. Thus, computational methods are preferred for bioprospecting molecules with potential repellent features. So far, *in silico* studies predicted and identified substances likely to interact with binding sites that are able to trigger a behavioral response.[49]

Many molecular docking approaches against OBPs have been attempted to analyze interactions with possible repellents. Vinay Gopal & Krishnan have found that the ligands camphor, carvacrol, oleic acid, and firmotox establish H-bond interactions

with 7 OBPs (PDB IDs: 3K1E, 1QWV, 1TUJ, 1OOF, 2ERB, 3R1O, OBP1)[50] from *Nilaparvatha lugens*. Dhivya and Manimegalai showed that di(2-Ethylhexyl)phthalate, beta amyrin, and alpha amyrin from the plant *Calotropis gigantea* are strongly bind with an OBP (PDB ID: 2L2C) from *Culex quinquefasciatus.*[34] Qadir and Arshad determined low interaction of the compounds bioneem and azadirachtin from *Azadirachta indica* with an *Anopheles gambiae* OBP (PDB ID: 3R1O).[51] However, Jayanthi *et al.,* employing Computational Reverse Chemical Ecology, screened and predicted active semiochemicals, namely Allyl isothiocyanate, E-2-Hexen-1-ol, E-4-Hexen-1-ol, and Z-3-Hexen-1-ol against *Plutella xylostella* Linn. OBP (PDB ID: 2WC5).[52] Thireou *et al*. filtered a library of 42 755 synthetic molecules based on i) the shape and chemical similarity to known plant-derived repellents, and ii) on the predicted similarity of the ligand's binding mode to the *Anopheles gambiae* Odorant Binding Protein 1 (*Agam*OBP1) relative to that of DEET and Icaridin to perform a prospective screening. Then, they tested 16 of these compounds *in vitro* in *Agam*OBP1. They found no correlation between repellence and OBP-binding strength. Nevertheless, a correlation between binding mode (the respective and stable position between OBP and ligand) and repellence was found.[53] On the other hand, da Costa *et al*. screened 1633 essential oil compounds based on the similarity to DEET binding. They analyzed the interaction with the OBP1 homodimer of *Anopheles gambiae*, and determined high-affinity with thymol acetate, 4-(4-methyl phenyl)-pentanal, thymyl isovalerate, and p-cymen-8-yl.[48] Nonetheless, a correlation between affinity and repellent activity is not established either. Besides, Portilla-Pulido *et al*. designed a repellent against *Aedes aegypti* (PDB ID: 3K1E) with effectivity as comparable to DEET using in silico simulations were complemented with *in vivo* repellent activity bioassays.[49]

Several studies related to docking with OBPs have been carried out. However, the affinity discriminates against the repellent activity. There are various shortcomings to studies on repellents, such as i) the analyses have not been previously calibrated; ii) so far they have only focused on one specific OBP. Hence, the vast majority of OBPs has not been analyzed yet to find broad-spectrum agents; iii) most of the experiments were dock-based building on the similarity of the compounds to DEET, which does not allow for finding new seeds; iv) the screening for interactions have only been used to select new agents from known libraries, and v) it has been imposible to establish a relationship between affinity and repellent activity. Therefore, it is utmost important to search for new scaffolds. Compounds other than DEET, terpenoids, and related substances have to be found in a way that allows us to broaden our search for compounds of different chemical nature and with desirable characteristics.

Quantitative Structure-Activity Relationships (QSAR) modeling is a good approach for drug discovery and has been widely used to search for new repellents. For instance, Katritzky *et al*. used molecular descriptors of a library of acylpiperidines calculated by CODESSA PRO software, aiming to model the relationship between mosquito repellency and the chemical structure of the compounds.[54] Paluch *et al*. analyzed a set of 12 sesquiterpenes through static-air bioassay, and aided by classic and quantum molecular descriptors, they developed a QSAR model for repellency.[55] Oliferenko *et al*. found a few highly active compounds as viable candidates of mosquito repellents using the *Aaeg*OBP1, based on molecular field topology analysis and scaffold hopping.[56] However, after performing a docking analysis, they did not find a correlation between activity and the compound docked on OBP. It should be noted that machine learning (ML) techniques based on odorant chemical descriptors also allow predicting new ligands for a given receptor.[57] Thus, Janairo *et al*. used 20 DFT

calculated descriptors to predict the repellency of 33 plant-derived compounds.[58] Kepchia *et al*. analyzed 1280 odorant molecules and identified active antagonists for the conserved Orco.[59] Recently, Caballero-Vidal *et al*. studied Lepidoptera and screened 3 million molecules that allowed them to find 11 novel agonists on *Spodoptera littoralis* OR through Support Vector Machine.[57] Several of the studies described above are based on congeneric data, that is, data from families of related compounds. So far, QSAR modeling has been performed several times. However, nobody has implemented QSAR models in expert software to discover new repellents. To this day, there is no free, cross-platform, and easy-to-use tool that facilitates to find leading compounds without the need for prior pure and theoretical knowledge of this topic.

1.7. Objectives and Structure of the Report

The present study aims to introduce a combination of molecular docking, QSAR modeling, and ML techniques implemented in user-friendly software that allows for the discovery of novel, potential repellents with suitable characteristics. Thus, we started with calculating the QuBiLS molecular descriptors (2D, [60] 3D,[61] and a fusion of both) of the sensilla of the mosquito *C. quinquefasciatus*. The quantitation of the interaction ability proceeded, by calculated with AutoDock Vina[62] of a repertoire of 13 OBPs from different species. These calculations allowed to generate classification, regression, and ensemble models against 50 compounds with known experimental repellency activity for each of the six sensilla of *C. quinquefasciatus* mosquito.[29] These exhaustive procedure endorsed ro find compounds able to induce a response in the sensilla. The models are implemented in an expert software named SiliS-PAPACS. Furthermore, a retrospective analysis is carried out with four different datasets of compounds, for that repellent activity has been previously reported, to calibrate the predictions made for the effects on the sensilla and to validate the software. Finally, a

prospective screening is performed using SiliS-PAPACS software to identify new IRs scaffolds by sifting different libraries of drug-like compounds and natural products.

2. EXPERIMENTAL PROCEDURES

2.1. Datasets

For the development, validation, and prospection of the regression and classification models, three different experiments (**EXP**) were carried out.

In the first experiment (**EXP1: Modelling)**, fifty compounds were extracted from the Liu *et al*'s. [29] research. For each compound its olfactory response has been recorded in spikes/s in each of the sensilla from *Culex quinquefasciatus*; see Supporting information **SI_1_Table 1**. Besides, the fifty IRs used in this study are given as Supporting information, folder **SI_2_Insect_OBPs_PDB**. Sensilla are of the long sharp trichoid (LST), short sharp trichoid (SST), short blunt trichoid I (SBT-I), short blunt trichoid II (SBT-II), short blunt trichoid-curved (SBT-C), and grooved peg (GP) types. However, LST and GP showed no significant response to the repellents tested in previous studies.[29] Thus, for the experiments, we used the following six types of sensilla: SBT-I type A and B (SBT-I-A and SBT-I-B), SBT-II type A and B (SBT-II-A and SBT-II-B), SST, and SST-C. A and B types differ in the spike amplitude's length, A producing the larger spike amplitude and B the smaller one.[29]

This experiment is divided into two parts. **EXP1A** includes QSAR Modeling using ML techniques, and molecular descriptors QuBiLS-MAS[60] and QuBiLS-MIDAS.[61] In contrast, **EXP1B** involves predictive modeling with the same dataset of fifty IRs however the response variable is the binding affinity with thirteen OBPs (**SI_1_Table 2**). These last were calculated through molecular docking calculations using the program AutoDock Vina (AV).[62] The PDBs' OBPs are as follows 1N8V,

1OW4, 1QWV, 2GTE, 2GVS, 2WC5, 3FIQ, 3K1E, 3N7H, 3OGN, 3PM2, 3R1O, and 3S0D (**SI_1_Table 2**). These are found in ten different insects: *Mamestra brassicae, Rhyparobia maderae, Antheraea polyphemus, Drosophila melanogaster, Schistocerca gregaria, Bombyx mori, Aedes aegypti, Anopheles gambiae, Culex quinquefasciatus,* and *Apis mellifera*, respectively; and the rodent *Rattus norvegicus*.

The **EXP2** involves four datasets for a retrospective screening. The first dataset, "A," is composed of 71 carboxamides found through QSAR and molecular docking experiments complemented with bioassays in which the Minimum Effective Dosage (MED) in μmol/cm^2 was measured against female *Aedes aegypti*.[56] The second one, "B," has 34 carboxamides the repellency of which was tested against the three most common breeds of cockroach: *Periplaneta americana*, *Blattella germanica*, and *Blatta orientalis.*[63] The third dataset, "C," has 34 essential oils and DEET, in total 35 compounds, whose repellency was tested on forearms of human volunteers against *A. gambiae* sensu stricto.[64] The fourth set, “D”, has 13 botanical sesquiterpenes assayed for spatial and contact repellency against *A. aegypti*.[55] These four datasets are provided as Supporting information, folder **SI_3_EXP2_Retrospective Study Datasets**.

The **EXP3** involves the Malaria Box[65], [66] dataset, and a library of 791 different Essential Oil Constituents (EOCs) for a prospective screening using the software SiliS PAPACS. Malaria Box collects 400 diverse chemotypes that include 200 compounds with drug-like properties and 200 probe-like confirmed blood-stage active antimalarial compounds.[66] This dataset is for the first time screened for drug repositioning to explore the repellent properties. On the other hand, the EOCs library comprises the main metabolites from plant extracts that might be related to repellency and the ideal properties of the IR; explored in a way of going back to the natural roots

of the pharmaceutical industry. Both datasets are available as Supporting Information, folder **SI_4_EXP3_Prospective Analysis**. The entire methodology applied in this work is presented in **Figure 1**. This figure summarizes all the steps followed to obtain new classification, regression, and repellency models based on QSAR and docking experiments, as well as the prospective procedure to find lead IR compounds.

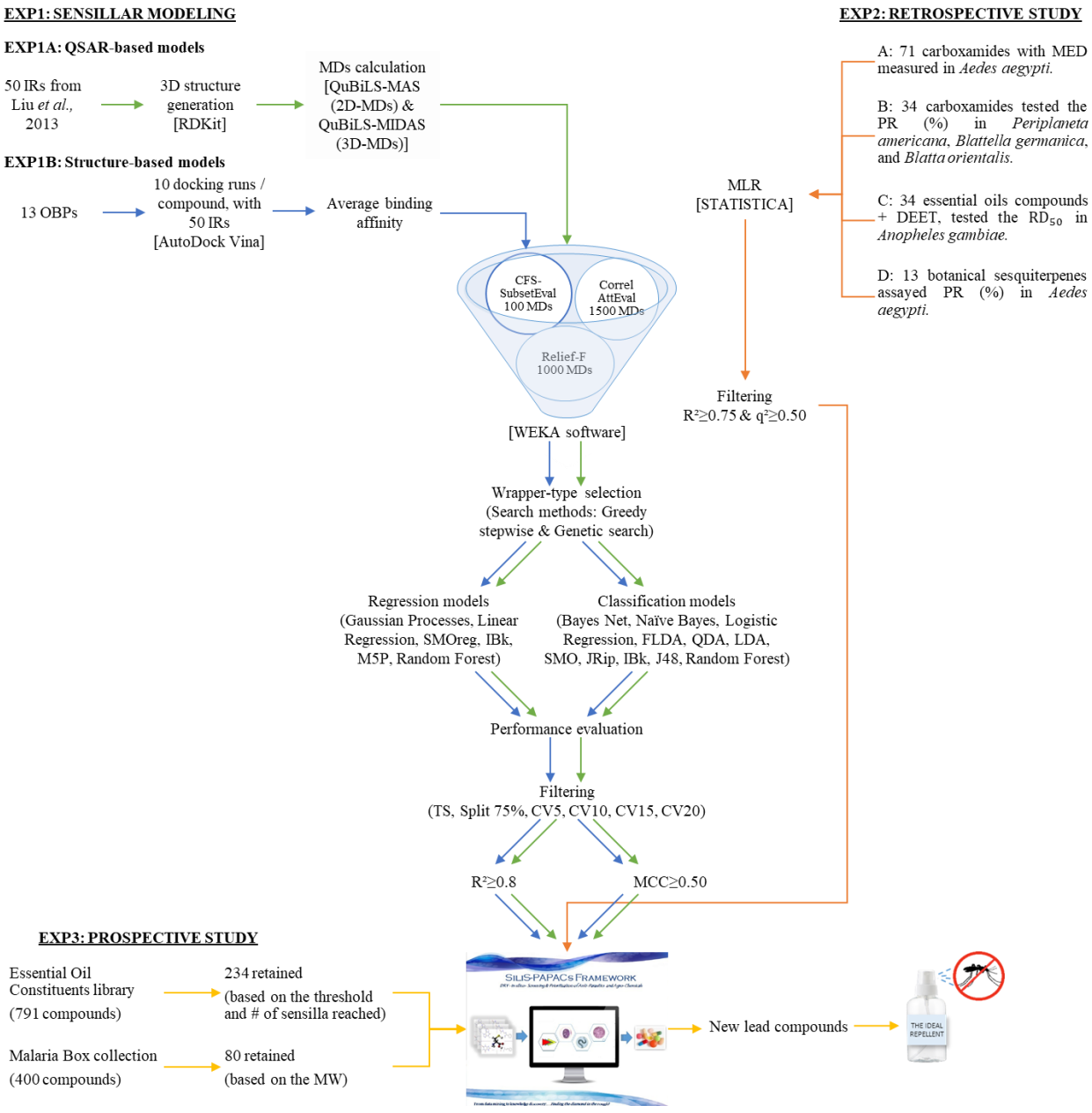

Figure 1. Comprehensive overview of the methods section. The software used is shown in brackets and the procedure specifications are shown in parentheses.

2.2. Structural coding of IRs

The tridimensional (3D) geometry for the structures was generated with the cheminformatics toolkit RDKit[67] (https://www.rdkit.org/) by distance geometry. The 3D conformations were optimized by using the Molecular Mechanic Force Field (MMFF94). SDF file is available in Supporting Information **SI_5_EXP1_3D-RDKit_MMFF_Repellents_liu(2013)**.

2.2.1. Molecular descriptors calculation

A number of 4463 two-dimensional (2D) molecular descriptors (MDs) were obtained from **the EXP1A** dataset by using QuBiLS-MAS software[60] (http://tomocomd.com/qubils-mas). These calculations are based on the bilinear, quadratic and linear algebraic forms.[60] These indexes have been used previously in different endpoints giving great predictive results.[68], [69] Unsupervised attribute selection was performed with IMMAN software[70] (http://mobiosd-hub.com/imman-soft/) to reduce the dimensionality in the dataset that was later processed in WEKA's Feature Selection module. The run parameters used were Shannon Entropy (SE) = 0.8 and Spearman correlation = 0.6. The MDs list was obtained as a Duplex's TXT file.

Additionally, a list of 2658 tridimensional (3D) MDs was obtained using the **EXP1A** dataset, as well, with QuBiLS-MIDAS software[61] (http://tomocomd.com/qubils-midas). The calculations are based on multi-linear or N-linear algebraic forms.[61] The run parameters were specified as Shannon Entropy (SE) = 0.8 and Spearman correlation = 0.5. This list was saved as a Ternary's TXT file.

In the case of the Duplex dataset, it was reduced to 3450 non-repeated MDs against the 50 IRs. In contrast, the Ternary one was filtered to different 2124 MDs. Each CSV dataset (duplex and ternary) was added as dependent variable the olfactory response from each of the six sensilla against the 50 IRs.[29] Thus, six datasets were

generated per MDs set type (Duplex and Ternary) for a total of twelve to be used as variables to predict the spikes/s response.

2.2.2. Molecular docking

The compounds were studied for their binding affinity to thirteen OBPs using the molecular docking program AV, which combines some advantages of knowledge-based potentials and empirical scoring functions. A grid-based protein-ligand interaction is used to speed up the score calculation since ligands are ranked based on this energy scoring function.[62]

For each of the PDB structures evaluated, a cube was established at their geometrical center in order to determine the docking site for the OBP structure. Ten runs were performed for each compound. All docking calculations included 20 number modes, an energy range of 1.5 kcal/mol, and exhaustiveness equal to 25. Finally, the best poses' average binding affinity was accepted as the binding affinity value for a particular complex.

Identification of Interacting Residues*:* Identifying residues (PDB: **3N7H** from *A. gambiae*) that interact on the binding site with the DEET model with the highest affinity value and the best pose was performed using Ligand Scout 3.11 (see **Table 1** and

Figure 2 and **Figure 3**).[71] Residue-ligand interactions were visualized with the program PyMol.[72]

Table 1. AutoDock Vina (AV) Affinities and RMSD Values Produced from Binding Best Model of DEET on OPB Structure (PDB: 3N7H), as well as Residues Interacting in the Binding Pocket (cutoff 4Å).

Co-crystalized ligand	DEET native (co-crystalized)	DEET model 15 (run 8)	DEET model 3 (run 41)	DEET model 1 (run 1)
AV affinity (kcal/mol)	-	-6.10 ± 0.00	-6.70 ± 0.00	-6.80 ± 0.00
RMSD (Å)	-	0.34 ± 0.00	0.42 ± 0.00	10 ± 0.00
Residues (contact at cutoff of 4Å)[a]				
Leu-15				+
Leu-19				+
Leu-58				+
Phe-59		+	+	+
Leu-73	+	+	+	
Glu-74		+	+	
Leu-76	+	+	+	+
His-77	+	+	+	
Leu-80	+	+	+	+
Met-84		+	+	
Ala-88	+	+	+	+
Met-89	+	+	+	
Met-91	+	+	+	
Gly-92	+	+	+	
Lys-93	+	+	+	
Leu-96’	+	+	+	
Arg-94	+	+	+	
Leu-110		+	+	
His-111		+	+	
Trp-114	+	+	+	
Trp-122				+
Phe-123				+
Leu-124				+
Val-125				+
Water-153’	+		+	
Water-350				+
Water-360	+	+		

[a]Residues from chain B are indicated with a prime (’) and in bold type. Interactions were assigned for atoms separated by ≤ 4Å.

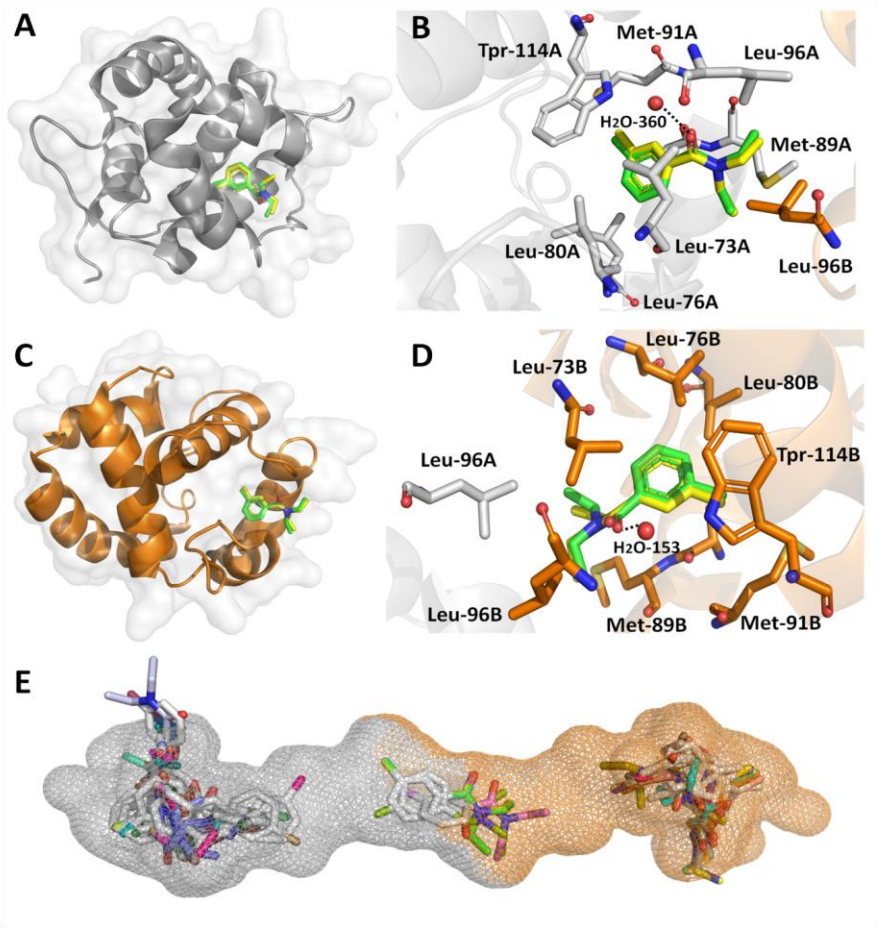

Figure 2. Representation for fifty re-docking runs to DEET into their respective binding sites on AgamOBP1 (PDB ID: 3N7H from A. gambiae). Crystallized conformation for each ligand is shown in green. The best re-docked pose is depicted in yellow for all complexes. A) 3N7H chain A complex-DEET model 15, run 8. B) Residues in the interaction 3N7H chain A complex-DEET model 15, run 8. C) 3N7H chain B complex-DEET model 3, run 41. D) Residues in the interaction 3N7H chain B complex-DEET model 3, run 41. E) DEET's re-docking run binding site is located at the center of a long hydrophobic tunnel (represented as a mesh) running through the dimer interface. Here, DEET molecule is bound to each subunit at a site located near the interface between the two monomers (center). However, the remaining area of each cavity is filled with other DEET poses (with marginal best AV affinities, see Table 2), which originally was occupied by PEG molecule that is used as a crystallization agent.

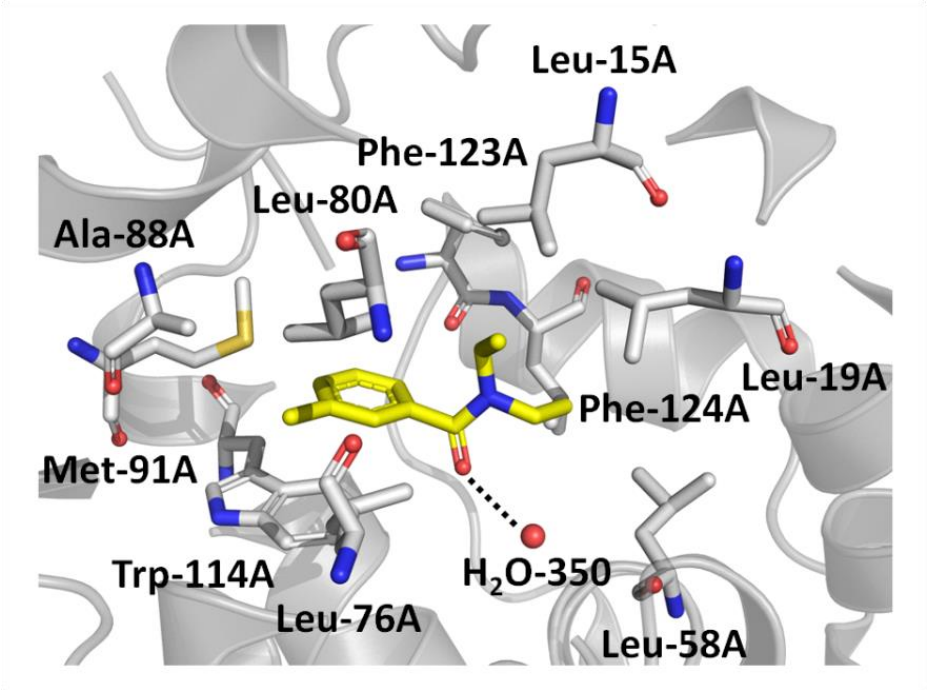

Figure 3. Schematic representation of the energetic binding site more favorably predicted for Autodock Vina (DEET model 1, run 1: -6.8 ± 0.0 kcal/mol, see **Figure 2E**). DEET (with carbon atoms in yellow) form one H-bond with a water molecule (red). The contacts between DEET and AgamOBP1 (3N7H) are dominated by non-polar van der Waals (vdW) interactions (see **Table 1**).

2.3. Modeling

2.3.1. Feature Selection

The selection attributes were performed with Waikato Environment for Knowledge Analysis (WEKA)[73] ML suite to filter the variables and obtain subsets for the prediction modeling in **EXP1A** and **EXP1B**. The workflow was reiterated for each data set of the six sensilla.

EXP1A started by i) applying the Correlation Attribute Evaluator with Ranker as search method and specifying 1500 as the number of variables to select. ii) Consecutively, this result was filtered again using the Relief-F method to retain the 1000 nearest features. iii) Finally, the Correlation-based Feature Selection (CFS) Sub

Evaluation algorithm was applied with Genetic Search and Greedy Stepwise search methods to retain up to 100 variables per subset. **EXP1B**, on the other hand, started directly with the Wrapper selection step.

Different subsets were generated with Wrapper-type selection.[74] In the case of Classification modeling, eleven ML techniques applied were Bayes Network (BN), Naïve Bayes (NB), Fisher's Linear Discriminant function (FLDA), Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA), Logistic regression function (Log), John Platt's Sequential Minimal Optimization (SMO, with Pearson Universal Kernel (PUK)) algorithm, Stochastic Gradient Descent (SGD), IBk (with 10 K-nearest neighbors and True cross-validation (CV)), JRip rule, and J48 and Random Forest (RF) trees. All classifiers' parameters described above were set as the default ones in WEKA.

Regarding Regression modeling, in the Wrapper-type selection, the Greedy Stepwise and Genetic Algorithm strategies were used as search methods. The ML methods were also used with the default parameters in WEKA software. Six classifiers were used in Regression modeling: Linear Regression (LR), RF, M5P tree, Gaussian Processes (GP), and Sequential Minimal Optimization (SMOreg) algorithm (both with Pearson Universal Kernel (PUK)), and IBk (with 10 K-nearest neighbors and True CV).

2.3.2. Classification Modeling

Classification models were developed to predict the spike amplitude of the extracellular record of ORN response potentials of the antennal trichoid sensilla to *C. quinquefasciatus* in two classes: POSITIVE (active) or NEGATIVE (inactive). The breakpoint for the sensilla are given in spikes/s: SBT-I-A = 13.2, SBT-I-B = 16.3, SBT-II-A = 20.4, SBT-II-B = 32.8, SST = 14.3, and SST-C = 13.5.[75] Values greater than or equal to the breakpoint will elicit a response in this specific sensillum and are represented as POSITIVE. Lower values represent certain actions occurring; however,

these are not enough to activate the sensillum response; these are classified as NEGATIVE.

For this purpose, WEKA's Classify module was used. The same eleven ML techniques were applied as well as in Wrappers selection to evaluate the performance based on 10-fold CV training. Then the meta-classifiers applied were Bagging, Vote, and Stacking with different combination rules for each. This process was applied to duplex and the ternary dataset, and from the best models of both, the MDs were used to create a new dataset that best combines the different descriptors. The evaluation considered the 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 statistical parameters.[76], [77]

It is important to remark that the dataset of the sensillum SBT-I-B evaluated in this report was unbalanced; that is, the dataset had notably more negative than positive cases. Once unbalance was detected, the Synthetic Minority Over-Sampling Technique (SMOTE)[78] was applied. SMOTE used different combinations of nearest neighbors (5 and 10), and the percentage of over-sampling was 110%.

For the development of **EXP1B**, precisely the same ML techniques and evaluation methods to develop models were applied that describe the interaction of different OBPs against the fifty IRs.

2.3.3. Regression Modeling

The modeling process was built on case-based learning[79] by applying the six ML classifiers described before each of the subsets generated through Wrappers. The meta-classifiers were also applied to each dataset: Additive Regression, Bagging, Stacking, and Vote with different combination rules. In each subset of sensilla, the retain of robust individual and ensemble models was based on 10-fold CV. Moreover,

the filtering was based on multi-criteria decision-making[80] to assess the performance, stability, the number of instances involved, and the diversity of ML methods applied. The predictive power of the models was evaluated based on the statistical parameters: correlation coefficient (R^2), Mean Absolute Error (MAE), and Root Mean Square Error (RMSE).

The procedure described above was applied to duplex as well to the ternary dataset and by using the MDs involved in the best models of both, it was created a new dataset that optimally combines the different descriptors. Furthermore, aiming to build a model that enhances the classification based on the predictions of the individual models obtained in each sensillum, an ensemble model was manually assembled. Selected subsets of the best individual and ensemble prediction models with 2D, 3D, as well as a fusion of both types of MDs, were obtained to improve the prediction of the spike amplitude of the extracellular record of ORN response potentials at the antennal trichoid sensilla to *C. quinquefasciatus* repellents of each sensillum against 50 different IRs.

Individual regression models were also obtained during **EXP1B** by applying the same ML techniques in which the variables that made up the models were a combination of the thirteen OBPs from different species against the sensillar response.

2.3.4. Applicability Domain

An analysis of the Applicability Domain (AD) was performed to understand the predictions' reliability with the models built.[81]–[83] A consensus-based decision among five estimation methods was considered, i.e., Range, Euclidean, Density, Manhattan, and Mahalanobis. The creiterion used in this work is that a consensus of at least 3 cases outside the AD methods' bounds reflects an unreliable prediction. The AD methods were applied with their default configurations and are available in the Ambit Discovery software.[84]

2.4. Retros-Prospective virtual screening

The models obtained in **EXP1A** and **EXP1B** were implemented in an expert system named SiliS-PAPACS that stands for DRY – *in **sili**co* – **S**creening & **P**rioritization of **A**nti-**P**arasitics and **A**gro-**C**hemicals **S**oftware, freely available at http://tomocomd.com/apps/papacs. The workflow of SiliS-PAPACS internally implements of supporting softwares, namely: RDKit,[67] QuBiLS-MAS[60] and QuBiLS-MIDAS,[61] Ambit Discovery,[84] AutoDock Vina,[62] Jmol,[85] and WEKA.[73]

Four different datasets[55], [56], [63], [64] were analyzed in **EXP2** to calibrate the prediction of repellent activity of the SiliS-PAPACS software. The potential to perform molecular docking by 50 IRs on a set of 13 OBPs was demonstrated so that the action on six different types of sensilla of the *Cx. quinquefasciatus* mosquito could described and classified. In this section, this study involves the IR experimental activity in four datasets, which include carboxamide derivatives and terpenoid-like scaffolds, considered the two most principal chemotypes with repellent action.[4], [55], [56], [64], [86]–[90]

Previous studies indicate that the OBPs mediate the first steps in the process of olfaction, including; uptake of odorants, transport through the sensillar lymph, and delivery to ORs in ORNs, triggering a sequence of biochemical events translated into behavioral responses.[45], [50], [91]–[93] These interactions are potentially attractive targets for QSAR analysis. OBP-derived calculations can be used as predictors for repellency or attraction of natural and synthetic semiochemicals[56]. In order to relate the values of changes of AV binding affinity of the IRs in four sets (for more details, see section 2.1 Datasets: **EXP2**) with activity repellent data,[55], [56], [63], [64] the

Multiple linear regression (MLR) analysis was applied using the program STATISTICA 8.0.[94]

Finally, with the software settings already adjusted, we proceeded to carry out **EXP3**, which involves prospective screening using the different libraries of drug-like compounds and natural products to find new seeds and lead compounds.

SiliS-PAPACS is the first tool that assists researchers in predicting the interaction of smell with the repellent activity of any given molecule. Additionally, internal cross-platform integration simplifies the time typically spent doing this type of research and allows finding leading compounds with structures totally different from those currently studied and with a broad spectrum.

3. RESULTS AND DISCUSSION

3.1. Performance of the Best Individual and Ensemble Models

Classification and Regression models were obtained for each of the six sensilla of *C. quinquefasciatus* mosquito as endpoints. Each model uses the 50 IRs as the independent variable (instances), and the standard dependent one for every case is the experimental electrophysiological value obtained in spikes/s of the olfactory response of the sensillum against the IRs[29] (**SI_1_Table 1**). In all the cases, WEKA software (v3.9.4) was used to generate the models. In **EXP1A**, the models involve a combination of a QuBiLS MDs matrix that followed a selection attribute process. In contrast, **EXP1B** uses a combination of the affinity values of 13 different OBPs against the 50 IRs obtained by using AV.

3.1.1. QSAR based models (EXP1A)

Classification models based on 2D, 3D, and a combination of both types of MDs (2/3D) were built. The MDs were calculated with QuBiLS-MAS[60] (2D) and

QuBiLS-MIDAS[61] (3D) software. Then, the MDs there were filtered through WEKA[73] and by using 11 different Wrapper techniques (described in the Methods section), and small subsets were generated. Around 12 classification models for each sensillum dataset that depend on the MDs of 50 molecules experimentally evaluated by Liu *et al.* [29] (**SI_1_Table 1**) was obtained. The models represent the action over each sensillum to discriminate them between a POSITIVE and NEGATIVE response. After development, these models were examined, and the ones with the best performance are presented in **Table 2**. As it can be observed, the number of QuBiLS-MAS 2D-MDs, QuBiLS-MIDAS 3D-MDs, and the fusion of them, are ranging from 3 to 9 MDs. A good model behavior during 10-fold CV with values of $0.612 \leq MCC \leq 1$, $80.597\% \leq Q \leq 100\%$, and $0.806 \leq Precision \leq 1$ is also shown. The best performances according to the MCC were obtained in sensillum SBT-II-B. The model C1_SBT-II-B (MCC = 1, Q = 100%, Precision = 1) involves 8 different 2D-MDs and uses the Bagging algorithm with Logistic function as classifier. Similarly, C2_SBT-II-B (MCC = 0.941, Q = 98%, Precision = 0.98) involves 6 descriptors as a result from the mix of 2D and 3D MDs; this uses the FLDA algorithm. In general, most of the ML techniques applied are the meta classifiers Vote and Bagging.

Table 2. Performance of the Classification models of **EXP1A** based on the statistical parameters obtained from WEKA through 10-fold CV training.

Sensillum	ID model	Technique	# of MDs	MCC	ROC Area	PRC Area	TP Rate	FP Rate	Q (%)	Precision
SBT-I-A	C1_ SBT-I-A	Vote using a combination of the average probability of the classifiers: LDA, SGD, and SMO	4	0.729	0.836	0.838	0.86	0.158	86	0.873
	C2_ SBT-I-A	Vote using a combination of the average probability of the classifiers: IBk, NB, and SMO	6	0.678	0.862	0.87	0.84	0.169	84	0.841
SBT-I-B	C1_ SBT-I-B	FLDA	9	0.612	0.839	0.846	0.806	0.194	80.597	0.806
	C2_ SBT-I-B	Bagging with J48 as classifier	3	0.612	0.839	0.815	0.806	0.195	80.597	0.806
SBT-II-A	C1_ SBT-II-A	Bagging with NB as classifier	5	0.683	0.87	0.862	0.84	0.175	86	0.847
	C2_ SBT-II-A	Vote using a combination of the average probability of the classifiers: QDA, NB, Log and SMO	5	0.645	0.894	0.904	0.82	0.198	82	0.831
SBT-II-B	C1_ SBT-II-B	Bagging with Log as classifier	8	1	1	1	1	0	100	1
	C2_ SBT-II-B	FLDA	6	0.941	1	1	0.98	0.071	98	0.98
SST	C1_SST	Stacking with BN as metaclassifier and the classifiers: BN, FLDA, Log, SGD, and RF	7	0.80	0.941	0.919	0.90	0.082	90	0.909
	C2_SST	Bagging with SGD as classifier	7	0.876	0.945	0.941	0.94	0.098	94	0.945

SST-C	C1_SST-C	FLDA	9	0.758	0.918	0.895	0.88	0.09	88	0.896
	C2_SST-C	Vote using a combination of the minimum probability of the classifiers: SGD, LDA, IBk, SMO, and Log	7	0.843	0.878	0.862	0.933	0.108	84	0.933

Regression models were developed by using the previously calculated MDs to quantitatively predict the olfactory response of the 50 IRs against each sensillum. Initially, around 60 individual models after Wrappers-selection in Weka workbench were obtained. The number of QuBiLS-MAS (2D) and MIDAS (3D) MDs range from 6 to 10. The most robust models are shown in **Table 3**. It can be observed that the models represent good behavior, for instance, $0.7173 \leq R \leq 0.9882$, $4.613 \leq MAE \leq 9.1113$, and $6.6334 \leq RMSE \leq 11.2657$. In addition, the best individual models regarding the R are from sensillum SBT-II-B. The one with the best R is I1_ SBT-II-B (R = 0.9882, MAE = 6.0311, and RMSE = 7.8035). It was built with the Bagging meta classifier algorithm with IBk as a classifier and involved eight 2D-MDs. I2_ SBT-II-B (R = 0.9873, MAE = 6.3461, and RMSE = 8.0678). The Additive Regression meta classifier was applied with IBk as a classifier to model 8 2D-MDs. From the compendium of models previously described, the meta classifier Additive Regression gave the best results.

Subsequently, from the best individual models, **Ensemble** models were built using the MDs involved in each. The same ML techniques as in the Individual Regression models were applied. **Table 3** shows the statistics of the best ensemble models for the six sensilla in CV 10. In this case, the number of MDs used goes from from 9 to 36 different 2D- and 3D-MDs. These integrative models' performance is $0.8968 \leq R^2 \leq 0.9962$, $0.734 \leq MAE \leq 3.7256$, and $1.6815 \leq RMSE \leq 6.7166$. Again, the best results are seen in sensillum SBT-II-B. E1_ SBT-II-B (R^2 = 0.9962, MAE = 2.9631, RMSE = 4.4936) as well as E2_ SBT-II-B (R^2 = 0.9961, MAE = 3.1799, RMSE = 4.5416) used the Vote meta classifier with 13 and 9 MDs, respectively. In general, LR is the most commonly used classifier that gave excellent results.

Table 3. Performance of the Best Individual and Ensemble Regression models of **EXP1A** based on the statistical parameters obtained from WEKA through 10-fold CV training.

Sensillum	Model Type	ID model	Technique	# of MDs	R^2	MAE	RMSE	RAE (%)	RRSE (%)
SBT-I-A	Individual	I1_ SBT-I-A	Vote using a combination of the maximum probability of the classifiers: GP, LR, and M5P tree	6	0.7558	7.9417	10.4025	60.286	67.2734
		I2_ SBT-I-A	Additive Regression with LR as classifier	6	0.7173	8.051	10.6235	61.1162	68.7026
	Ensemble	E1_ SBT-I-A	Additive Regression with LR as classifier	12	0.8968	3.6618	6.7107	27.7969	43.3985
		E2_ SBT-I-A	Vote using a combination of the maximum probability of the classifiers: LR, SMOreg and M5P tree	12	0.8969	3.7256	6.7166	28.2813	43.4365
SBT-I-B	Individual	I1_ SBT-I-B	Vote using a combination of the average probability of the classifiers: GP, SMOreg and IBk	6	0.9213	8.2697	10.7981	53.1897	40.6238
		I2_ SBT-I-B	Bagging with IBk as classifier	10	0.9174	9.1113	11.2657	58.6026	42.3829
	Ensemble	E1_ SBT-I-B	M5P	12	0.9889	1.8062	3.8634	11.617	14.5348
		E2_ SBT-I-B	LR	15	0.9889	1.8062	3.8634	11.617	14.5348
SBT-II-A	Individual	I1_ SBT-II-A	Vote using a combination of the	10	0.8569	5.9322	7.7384	67.988	59.4346

			maximum probability of the classifiers: GP, LR, and IBk						
	Ensemble	I2_ SBT-II-A	LR	10	0.8563	4.9871	6.6334	57.1569	50.9476
		E1_ SBT-II-A	LR	9	0.9633	1.9101	3.4367	21.8909	26.3952
		E2_ SBT-II-A	LR	17	0.9628	1.8602	3.4628	21.3198	26.5957
SBT-II-B	Individual	I1_SBT-II-B	Bagging with IBk as classifier	8	0.9882	6.0311	7.8035	15.5033	15.33
		I2_ SBT-II-B	Additive Regression with IBk as classifier	8	0.9873	6.3461	8.0678	16.313	15.8492
	Ensemble	E1_ SBT-II-B	Vote using a combination of the maximum probability of the classifiers: LR, SMOreg, and M5P	13	0.9962	2.9631	4.4936	7.6169	8.8276
		E2_ SBT-II-B	Vote using a combination of the maximum probability of the classifiers: LR, M5P, and SMOreg	9	0.9961	3.1799	4.5416	8.174	8.9219
SST	Individual	I1_SST	Additive Regression with IBk as classifier	8	0.8942	5.2584	6.7424	47.7345	44.2123
		I2_SST	Vote using a combination of the average probability of the classifiers: LR, IBk, and RF	8	0.8937	4.613	6.9977	41.8753	45.886
	Ensemble	E1_SST	LR	10	0.9837	1.1587	2.7192	10.5179	17.8304
		E2_SST	LR	14	0.9834	1.1925	2.6955	10.8251	17.6756

SST-C	Individual	I1_SST-C	Additive Regression with SMOreg as classifier	7	0.9133	6.1301	7.5744	44.0349	43.0061
		I2_SST-C	Additive Regression with GP as classifier	7	0.8933	5.3679	7.8914	38.56	44.8062
	Ensemble	E1_SST-C	LR	36	0.9953	0.734	1.6815	5.2726	9.547
		E2_SST-C	Vote using a combination of the maximum probability of the classifiers: LR, SMOreg, and M5P	9	0.9953	0.7427	1.6894	5.335	9.592

Regression and classification models allowed us to obtain a measure of performance in predicting the experimentally calculated olfactory responses based on MDs. The performance of these models and others implemented in the SiliS-PAPACS software is presented in **SI_1_Table 3** and **SI_1_Table 4**. However, all the models' information is presented as Supporting information in the folder **SI_6_EXP1A (QSAR-based models)**. This folder includes the starting SDF file, the ARFF and CSV output files obtained from the Weka software, a DOCX file presenting the description of the model built, a output file in XLSX format containing the statistical parameters information and predictions from the Weka training, and the MODEL file.

3.1.2. Structure-based models (EXP1B)

Once the affinity values of docking calculations through AV[62] using a panel of 13 different OBPs were obtained, these were used as variables to build classification and regression models to predict the response of the trichoid sensilla against the 50 IRs.[29]

Several 48 **Classification** models were developed through Weka software[70] through Wrappers selection, applying the ML techniques described in the Methods section. These models and classification QSAR-based models predict a POSITIVE or NEGATIVE response of the OBPs against 50 IRs in each of the six sensilla. The best models, that were implemented in SiliS-PAPACS software, are presented in **Table 4**. The number of OBPs used as variables varied from 2 to 6. The performance in 10-fold CV is $0.403 \leq MCC \leq 0.941$, $72\% \leq Q \leq 98\%$, and $0.74 \leq Precision \leq 0.98$. The models with the best behavior regarding the MCC values are seen in sensillum SBT-II-B. Model C1_SBT-II-B (MCC = 0.941, Q = 98%, Precision = 0.98) uses 6 different OBPs (OBPs' PDB ID: 1N8V, 1QWV, 2WC5, 3K1E, 3OGN, and 3R1O) and was trained with the Logistic regression function algorithm. Likewise, the C2_SBT-II-B model (MCC =

0.882, Q = 96%, Precision = 0.962) involves 4 different OBPs (OBPs' PDB ID: 1QWV, 3K1E, 3PM2, and 3S0D) and uses the IBk algorithm.

Table 4. Statistical parameters of the Classification models of EXP1B obtained from WEKA through 10-fold CV training.

Sensillum	ID model	Technique	# of OBPs	MCC	ROC Area	PRC Area	TP Rate	FP Rate	Q (%)	Precision
SBT-I-A	C1_ SBT-I-A	IBk	2	0.485	0.663	0.664	0.74	0.286	74	0.755
	C2_ SBT-I-A	Vote using a combination of the maximum probability of the classifiers: JRip, Log, and RF	2	0.475	0.728	0.721	0.74	0.267	74	0.74
SBT-I-B	C1_ SBT-I-B	Vote using a combination of the average probability of the classifiers: IBk, JRip, and RF	2	0.58	0.759	0.775	0.82	0.25	82	0.818
	C2_ SBT-I-B	FLDA	4	0.403	0.71	0.749	0.72	0.297	72	0.743
SBT-II-A	C1_ SBT-II-A	Stacking with QDA as metaclassifier and the classifiers: J48, JRip, and IBk	3	0.682	0.821	0.766	0.84	0.156	84	0.843
	C2_ SBT-II-A	Bagging with LDA as classifier	5	0.645	0.816	0.78	0.82	0.198	78	0.831
SBT-II-B	C1_ SBT-II-B	Log	6	0.941	0.944	0.962	0.98	0.071	98	0.98
	C2_ SBT-II-B	IBk	4	0.882	0.938	0.951	0.96	0.142	96	0.962
SST	C1_SST	J48	4	0.753	0.838	0.828	0.88	0.114	88	0.885
	C2_SST	RF	4	0.745	0.885	0.878	0.88	0.135	88	0.88

SST-C	C1_SST-C	Vote using a combination of the average probability of the classifiers: FLDA, Log, RF, and IBk	2	0.73	0.853	0.879	0.88	0.204	88	0.885
	C2_SST-C	IBk	6	0.587	0.8	0.767	0.8	0.189	80	0.818

Regression models allowed to predict a quantitative response in spikes/s in each sensilla involving multiple combinations of the 13 OBPs versus the 50 IRs. The models were developed through Wrappers in Weka software by applying the ML techniques previously described. The individual regression models with the best performance are presented in **Table 5**. As can be observed, the models involve from 2 to a maximum of 8 OBPs. The performance in 10-fold CV is $0.522 \leq R^2 \leq 0.9204$, $5.09 \leq MAE \leq 11.3597$, and $6.9172 \leq RMSE \leq 19.8615$. The sensillum with the best models regarding the R coefficient is SBT-II-B. For instance, I1_ SBT-II-B (R = 0.9204, MAE = 8.8472, and RMSE = 19.8615). It was built with the Bagging meta classifier algorithm with IBk as a classifier and involved 7 OBPs (OBPs' PDB ID: 1N8V, 1QWV, 2GTE, 3K1E, 3N7H, 3PM2, and 3S0D). Similarly, I2_ SBT-II-B (R^2 = 0.9201, MAE = 8.9845, and RMSE = 19.8056) also uses Bagging with IBk as classifier involving 5 OBPs (OBPs' PDB ID: 1N8V, 1QWV, 2GTE, 3K1E, and 3S0D).

Table 5. Statistical parameters of the Individual Regression models of EXP1B obtained from WEKA through 10-fold CV training.

Sensillum	ID model	Technique	# of OBPs	R^2	MAE	RMSE	RAE (%)	RRSE (%)
SBT-I-A	I1_ SBT-I-A	Additive Regression with RF as classifier	3	0.5581	8.9711	12.9976	68.1006	84.0563
	I2_ SBT-I-A	Additive Regression with SMOreg as classifier	4	0.522	11.3597	13.4138	86.2325	86.7476
SBT-I-B	I1_ SBT-I-B	SMOreg	5	0.9192	8.9736	12.0045	60.6631	44.5853
	I2_ SBT-I-B	Additive Regression with GP as classifier	3	0.9154	8.9247	11.1321	60.3328	41.345
SBT-II-A	I1_ SBT-II-A	Bagging with IBk as classifier	8	0.6165	7.0953	10.1934	81.3181	78.2899
	I2_ SBT-II-A	IBk	7	0.5616	5.09	6.9172	77.4194	82.5296
SBT-II-B	I1_SBT-II-B	Bagging with IBk as classifier	7	0.9204	8.8472	19.8615	22.7422	39.0179
	I2_ SBT-II-B	Bagging with IBk as classifier	5	0.9201	8.9845	19.8056	23.0952	38.9083
SST	I1_SST	Additive Regression with IBk as classifier	3	0.6656	7.9407	11.451	72.0838	75.0879
	I2_SST	RF	2	0.6484	8.8276	11.3854	80.1345	74.6578
SST-C	I1_SST-C	Additive Regression with IBk as classifier	4	0.6507	9.8954	13.3235	71.0832	75.6485
	I2_SST-C	Vote using a combination of the maximum probability of the classifiers: SMOreg and IBk tree	2	0.6458	9.9363	13.9128	71.3766	78.9941

Outliers: SBT-I-B: (-)-α-pinene, S-(-)-Perillaldehyde, Citronellal, and (-)-Linalool

In general, training with meta classifiers gave better results than simple classifiers. In all the models, the sensillum SBT-II-B showed the best performance, so its predictions with different molecules are thought to be better as well. Regarding the use of OBPs, an incidence matrix was made (presented in file **SI_1. Table 5**) that shows the variables that appear most frequently within the combinations of both classification and regression models in EXP1B. The most repetitive OBPs were found to be 11 times 2GTE and 3S0D, and 10 times 2GTE and 3S0D.

3.2. SiliS-PAPACS Software for Repellent Prediction

All the models developed, both for classification and regression, were implemented internally in the SiliS-PAPACS software. SiliS-PAPACS software has a user-friendly desktop interface (**Figure 4**). The steps to use the software involves six stages: i) selecting compounds from an SDF or MOL external file. ii) Then, it is generated the 3D structure with the internally implemented RDKit. If the molecules to analyze have been previously optimized, pass directly to the next step. iii) Choose the QSAR based regression and classification models to analyze and the methods to evaluate the AD. iv) Select the prediction models based on the structure and specify (optional) a time out for the docking (as default 10 minutes). v) Choose the method of clustering to group the molecules by similarity: K-means (non-hierarchical) or tree joining (hierarchical). And vi) Indicate the folder to save the results.

Once the screening is completed, a new interface is displayed, and a folder with the output information is automatically generated. The test sets used in the present report are provided as a sample dataset.

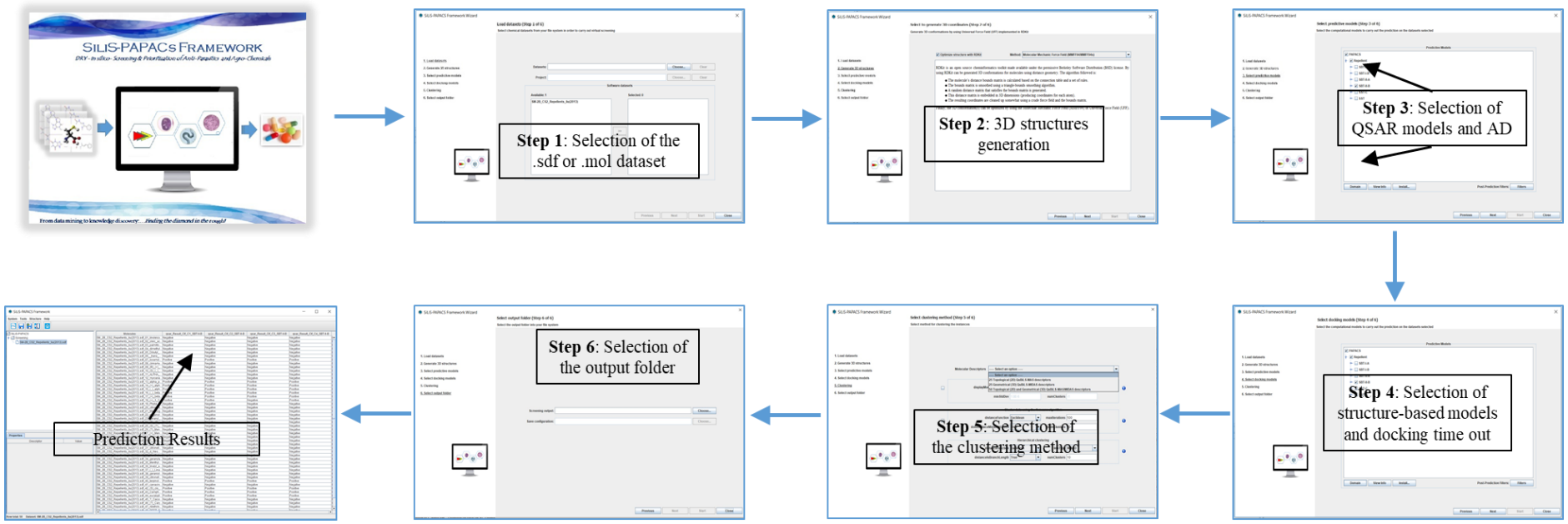

Figure 4. Screenshots of the interfaces of the SiliS-PAPACS software. Step 1: loading of external SDF or MOL file. Step 2: generation of the 3D structure (optional). Step 3: selection of QSAR-based models to evaluate and the AD. Step 4: selection of the structure-based models and the docking time out. Step 5: choosing the clustering method. Step 6: choosing the output folder

3.3. Results of the Retrospective Virtual Screening

The best models obtained based on the R^2 and q^2, together with their statistical parameters for each learning dataset, are given below:

$$\boldsymbol{Log(MED)} = -3.69(\pm 0.66) - 2.40(\pm 0.31) \times \boldsymbol{2GVS} + 0.27(\pm 0.07) \times \boldsymbol{3FIQ} + 0.28(\pm 0.15) \times \boldsymbol{3R1O} + 0.88(\pm 0.31) \times \boldsymbol{1QWV} + 0.34(\pm 0.13) \times \boldsymbol{3K1E} - 0.43(\pm 0.16) \times \boldsymbol{1OW4} \quad \textbf{(1)}$$

N= 61 R^2= 0.75 q^2= 0.67 s= 0.42 q^2_{boot} = 0.64 a(R^2) = 0.075 F(6,54)= 26.4 p< 0.0001

$$\boldsymbol{PR}(\%)_{B.\ germanica} = 103(\pm 32.1) - 30.4(\pm 3.73) \times \boldsymbol{3FIQ} + 54.4(\pm 10.7) \times \boldsymbol{1N8V} - 33.2(\pm 10.9) \times \boldsymbol{3OGN} + 48.8(\pm 11.7) \times \boldsymbol{3R1O} - 41.0(\pm 14.8) \times \boldsymbol{3PM2} \quad \textbf{(2a)}$$

N = 26 R^2 = 0.81 q^2 = 0.69 s = 11.7 F(5,20) = 16.9 p < 0.0001

$$\boldsymbol{PR}(\%)_{P.\ americana} = 241(\pm 35.9) + 45.9(\pm 9.23) \times \boldsymbol{2GTE} - 57.3(\pm 11.2) \times \boldsymbol{3OGN} + 75.6(\pm 13.5) \times \boldsymbol{2GVS} - 53.6(\pm 11.8) \times \boldsymbol{2WC5} + 72.2(\pm 14.9) \times \boldsymbol{3PM2} - 28.0(\pm 11.4) \times \boldsymbol{3R1O} \quad \textbf{(2b)}$$

N = 30 R^2 = 0.77 q^2 = 0.54 s = 11.5 F(6,23) = 12.8 p < 0.0001

$$\boldsymbol{RD_{50}} = -0.007(\pm 0.002) - 0.002(\pm 0.0006) \times \boldsymbol{2GTE} + 0.002(\pm 0.0004) \times \boldsymbol{3FIQ} - 0.001(\pm 0.0005) \times \boldsymbol{3R1O} \quad \textbf{(3)}$$

N = 30 R^2 = 0.75 $q2$ = 0.60 s = 0.001 F(4,25) = 18.6 p < 0.0001

$$\boldsymbol{Log(PR_{60min})} = 2.43(\pm 0.13) + 0.17(\pm 0.02) \times \boldsymbol{3R1O} - 0.14(\pm 0.02) \times \boldsymbol{3PM2} + 0.01(\pm 0.01) \times \boldsymbol{3S0D} \quad \textbf{(4a)}$$

N = 11 R^2 = 0.93 q^2 = 0.81 s = 0.02 F(3,7) = 29.7 p < 0.0001

$$\boldsymbol{Log(PR_{120min})} = 1.32(\pm 0.20) - 0.11(\pm 0.01) \times \boldsymbol{3FIQ} + 0.06(\pm 0.02) \times \boldsymbol{3N7H} - 0.03(\pm 0.024) \times \boldsymbol{3S0D} \quad \textbf{(4b)}$$

N = 11 R^2 = 0.91 q^2 = 0.78 s = 0.03 F(3,7) = 25.3 p < 0.0001

Where N is the number of cases, R^2 is the squared correlation coefficient, q^2 is the determination coefficient of the LOO cross-validation procedure, *s* is the standard deviation of the regression, and F is the Fisher ratio at the 95.0% confidence level. There are modeled the Minimum Effective Dose (MED) necessary to trigger a favorable response, the Percentage of Repellency (PR (%)), and the exposure concentration producing a 50% respiratory rate decrease (RD_{50}) [95]; and.The chemical SDF files for each dataset in this **EXP2** are given as **SI_3_Retrospective Study Datasets**. All results, including statistics, experimental and calculated values, as well as their linear relationships, outliers, and structure representation for these QSAR studies can be seen in the Supporting Information file **SI_1_EXP2_(A-D)**.

As can be seen, the statistical parameters obtained from the analysis of MLR follow a significant linear trend in all four cases, with R^2 (q^2) of 0.75 (0.67), 0.81 (0.69), 0.77 (0.54), 0.75 (0.60), 0.93 (0.81), 0.91 (0.78), respectively for all of the four cases studied: 71 carboxamide derivatives used by Oliferenko *et al*.,[56] 34 carboxamides assessed by Gaudin *et al*.,[63] 34 EOCs assessed by Omolo *et al*.,[64] and 12 sesquiterpene compounds of plant derivatives used by Paluch *et al*.,[55]. This indicates that the binding affinity predicted by AV for the different OBPs is linked to the repellent activity of the existing chemotypes in the group of compounds studied.

It is important to note that the OBPs **3R1O** (*A. gambiae*), **3FIQ** (*R. norvegicus*), and **3PM2** (*A. gambiae*) were identified in that order (they appeared 5, 4, and 3 times, respectively) as the most significant in all the obtained models (see **Table 6**). These proteins are classical and C-plus class OBPs, and they are found in large concentrations in the lymph fluid which surrounds the olfactory dendrites, indicating an important role in the process of olfaction.[23], [96] In fact, it is important to remark that all the OBPs in the study appear at least once in the obtained models, and there is no model in which just one OBP alone explains the repellent activity, which indicates that a set of various OBPs (a diverse repertoire) is necessary for an accurate description of the repellent activity.

Table 6. Frequency of the Appearance of the Variables (OBP-based Binding Affinities) in Regression and Classification Models.

OBPs	PLR[a]	LDA[b]	MLR[c]	Total
3FIQ	4	3	4	11
3S0D	3	4	2	9
1QWV	5	3	1	9
3R1O	2	2	5	9
2GTE	3	3	2	8
3PM2	3	1	3	7
1OW4	3	2	1	6
1N8V	3	1	1	5
2GVS	0	1	2	3
3K1E	0	1	1	2
3OGN	0	1	1	2
3N7H	1	0	1	2
2WC5	0	0	1	1

[a,b]Piecewise non-linear estimation (PLR) and linear discriminant analysis (LDA) equations for prediction of the olfactory response of ORNs of IRs in EXP2, respectively. [c]Multiple linear regression (MLR) based models for repellent activities description of molecules in EXP2.

Recently, Oliferenko *et al.*[56] did not obtain a good correlation between the values of binding affinity predicted by the Glide program with just one OBP (**3K1E**) with the repellent activity data. That is, in this study, a group of compounds that were least active in terms of MED was predicted to have similarly high docking scores that

compound with higher repellent activity. However, for the last class of compounds (seven times more potent), a high correlation coefficient of 0.92 suggests that they bind strongly to **3K1E**, which is linearly related to acting as a strong IR. Although this result reflects a direct link between the binding of **3K1E** and repellency, in fact, it is only a necessary but not a sufficient condition because it does not have a general relationship because, in the olfactory cascade, OBP is just one element in insect behavioral responses and in practice repertories of OBPs are present. For that reason, we propose to use a structurally diverse OBP repertoire and scale their affinities with ORN responses and repellent activities [56] because compound(s) with high affinity(ies) with OBP(s) may be slightly active (or completely inactive) on the second step of the ORs/ORNs transduction.

On the other hand, if we compare the results obtained by our model (Eq. **1**) (R^2 = 0.75 and q^2 = 0.67) with the reported recently by Oliferenko *et al.*[56] using Molecular Field Topological Analysis (MFTA) (R^2 = 0.96 and $q_{10\%}^2$ = 0.80), it can be seen that our results were lower, possibly because of the structural descriptors used by the authors in their studies (MFTA descriptors) are a direct quantification of molecule structures. That is, in **Eq. 1**, we used the AV affinities with a panel of 13 OBPs as variables, which are an "indirect" quantification of chemical structures because the binding affinities are docking scores from interactions of the complexes.[97], [98] However, the behavior of our approach in external predictions (see **SI_1_EXP2_A4** for observed and prediction values and **SI_1_EXP2_A6** for molecular structures) was similar-to-superior (R^2 = 0.23 for **Eq. 1**) than MFTA method (R^2 = 0.11 for MFTA model).[56]

Although the quantitative relationships of structure and repellency[18], [56], [105]–[109], [86], [90], [99]–[104] have been widely discussed in the research, only a small amount of quantitative data has been obtained; mainly carboxamides type

chemicals. The data used here with 71 compounds is the most representative (and biggest) and structurally diverse with this kind of scaffold from all these studies.[56] The second data of carboxamide selected by us is the only collection evaluated against cockroach[63] because all other studies mainly use mosquito species. However, this data is used for the first time here for the QSAR study, and thus it will not permit us to make a direct comparison. The other more explored chemotype as IRs are EOCs, where Omolo *et al*.'s[64] study is the biggest data reported so far, but also it is used here for the first time. Finally, to evaluate our approach for plant-related compounds in a comparative study, we selected a rather small data[55], [103] which is more structurally diverse than others previously used in QSAR studies, but that includes only α- and β-pinene terpenoid derivatives.[90], [107], [109]

In the last dataset in **EXP2**, the obtained models **Eqs. 4a** and **4b** are similar to those reported by Paluch *et al*.[55] with the same group of compounds and activities, performed models with several topological indices and posteriorly to carry out VS of a sesquiterpenes library. As compared to these results,[103] one can see that those results described here are similar-to-better. For instance, Garcia-Domenech *et al.*'s results are R^2 (q^2) = 0.88 (0.81) and 0.87 (0.77) *versus* R^2 (q^2) = 0.93 (0.81) and 0.91 (0.78) obtained by us at 60 and 120 min, respectively.

Finally, the adequate statistical quality of the QSARs built for these structurally diverse datasets attest its exploratory power and, with a good degree of accuracy (inside of DA,[110], [111] see next section **EXP3**), its capability to be used for VS purposes.

3.4. Results of the Prospective Virtual Screening to find new lead compounds

SiliS-PAPACS software was used for prospective VS of two sets of compounds, the first one is an Essential Oil Constituents library, and the second one is Malaria Box.[66] VS technologies have largely enhanced the impact of computational chemistry

within the lead discovery process and are now one of the computational tools used to filter out unwanted compounds from chemical libraries. In conjunction with High-throughput screening (HTS) technology, the VS has become the main tool for identifying leads, playing a predominant role in drug research.[112]–[117]

Many plant-derived products have been evaluated for their toxic properties against several insect species, especially the EOCs, which have been experimentally evaluated for decades, searching for repellent activity. Currently, several EOCs possess fumigant or contact toxicity,[118]–[121] repellent,[18], [87], [122] and antifeedant activity[123], [124] as well as development and growth inhibitory activity.[125], [126]

Considering that the most critical interest of any mathematical model developed is the use of *in silico* searching of new hits, VS's final experiment was carried out. In this section, the models previously developed to classify and predict the ORN responses as well as estimate the repellent activity in insects (RD_{50} and MED) will be used in an integrated way to computationally screen a diverse set of EOCs compiled by us from literature, as well as the Malaria box library. The chemical structure of these compounds and their bibliographic references are shown as Supplementary Material **SI_1_EXP3_(A & B)**. Besides, the SDF files of these compounds are given as **SI_4_EXP3**.

For the selection of the EOCs, the following criteria (rules) were used: 1) a compound will remain in the hit list if the predicted olfactory activity is higher than the threshold on 4-6 sensilla (equivalent to activating 4-6 ORNs) using LDA-based QSAR models of each sensillum, 2) if the predicted olfactory activity exceeds the threshold for only 1-3 sensilla, the compounds will only be selected as candidates if at least in one of these types of sensilla, the quantitative spike value predicted by the PLR model is similar to or higher than the compound that caused a strong stimulation at an

experimental level, 3) compounds that have activation above threshold in just one sensillum will be selected only if the repellent activity of the MLR models is potent or greater than the values of 8.9×10^{-5} mg/cm^2 and 0.052 μmol/cm^2 reported as potent by Oliferenko *et al.*[56] and Omolo *et al.*[64]

It is rather relevant to highlight that the predictions were only taken into account if the compound was within the AD of the models, for which we used a criterion consensus based on three different AD methods using the Ambit Discovery software (http://ambit.sourceforge.net). The AD of the QSAR model is "the range within which it *tolerates* a new molecule[127]–[129] This criterion is fundamental because one of the main aims of the present report was to develop models for predicting repellent activity at the early stages of drug discovery. Consequently, one may not mean *extrapolating* the use of these models to other kinds of chemical classes, making uncertain predictions in conditions very different from those fixed to calibrate the model. The majority of EOCs evaluated *in silico* fall within this area, which ensures great reliability for predicting these kinds of leads used in the VS.

The selected compounds using the three criteria above were 243 (243/791 = 30.7%), which reveals a very diverse group of chemicals; including many of them reported in the literature as IRs[64], [87], [89], [103], [107]. This shows the ability of the computational workflow to identify EOCs with repellent activity(ies) correctly. These compounds represent different structural types of terpenoids, sesquiterpenoids, monoterpenes acyclic, monocyclic and bicyclic, and diterpenes. Many studies show that the strong repellent activity of EOCs appears to be associated with the presence of these types of metabolites.[55], [87] Functional groups present in these compounds, such as hydroxyl, ketone, lactones, among others, are biologically active as IRs.[90], [130] Furthermore, computational approaches revealed that the repellent-receptor interactions

are most likely related to the electrophilic interactions.[90] In fact, QSAR models developed by Wang *et al.* [90] showed that MDS such as dipole moment and boiling point is closely associated with the repellent activity of terpenoid compounds. The first (dipole moment) could be regarding lipophilicity or specific electrostatic interactions with the receptor, whereas the latter (boiling point) can determine the duration of contact time with the olfactory chemosensilla of mosquitoes.[90]

It is important to note that the sensilla SST, SST-C, SBT-I-A, SBT-I-B, SBT-II-A, and SBT-II-B recovered 64.0, 59.0, 53.0, 54.0, 36.0, and 35.0%, respectively, of the compounds selected from screening (243 in total). Likewise, four compounds showed olfactory activity above the threshold in 6 sensilla, 12 in 5 sensilla, and 84 in 4 sensilla. In general, only 12.0% of the compounds in the dataset showed no activity on any sensilla. In total, only 16 EOCs shown significant activation of 5 or 6 sensilla. The majority of these EOCs are well-known IRs or are structurally rather-to-very similar to substances with proven insect repellent activity.

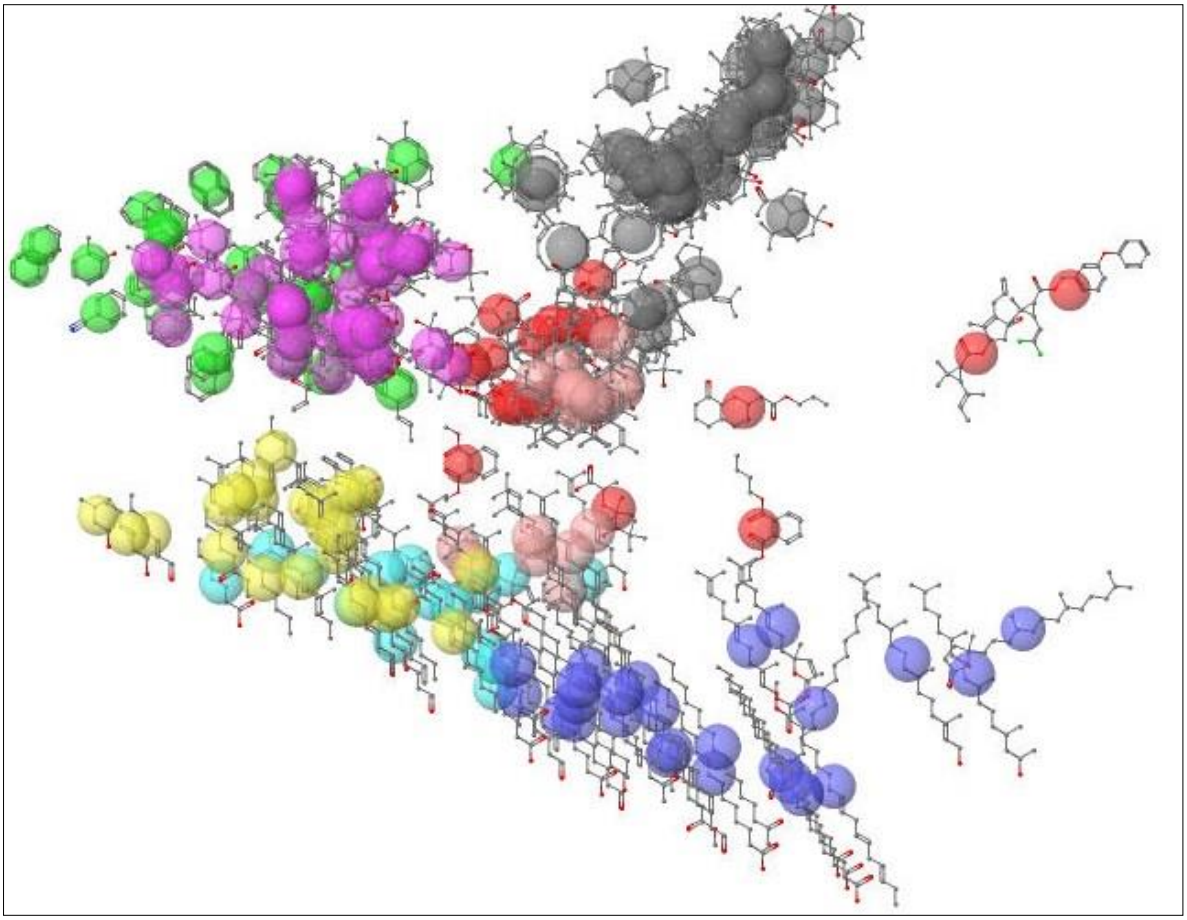

Figure 5. Clustering in 8 groups of the 234 EOCs selected from VS as potential IRs using the CheS-Mapper program.[131] These EOCs were compared and grouped with respect to the 50 IRs reported by Liu et al.[29] used in this study as an internal reference. The colors indicate compounds grouped by structural similarity. Cluster 1 (Purple) has 23 compounds, cluster 2 (red) presents 13 EOCs, cluster 3 with 26 compounds is green, cluster 4 (cyan) has 20 compounds, magenta represents the 52 compounds in cluster 5, cluster 6 (pink) has 13 chemicals, yellow highlights the 23 compounds in cluster 7 and cluster 8 (gray) is comprised of 64 compounds.

Furthermore, using the Chem*S*-Mapper program,[131] the 234 compounds selected by VS were compared and grouped with respect to the 50 compounds reported by Liu *et al.*[29] used in this study as an *internal reference* (queries to get the structural diversity of new hits) to verify to what extent these compounds are similar to those

reported in the literature and whether they represent new chemotypes. **Figure 5** shows the 234 compounds selected by VS, grouped by simple *k*-mean Cluster Analysis (CA).[131] As we can see, the EOCs were clustered into eight groups. The compounds in each 8 groups are resumed as 8 .sdf files in Supporting Information **SI_4_EXP3_234_Cluster**.

Here, it is essential to note that the groups (1-8 clusters) bring together a diverse set of promising chemicals capable of changing insects' behavior, some of them similar to those reported in the literature as IRs. In general, the clusters 4, 5, and 8 gather the compounds with theoretically good olfactory activity in 4, 5, and 6 sensilla (4, 12, and 84 compounds, respectively, for a total of 100 EOCs) and the best values of RD_{50} and MED (1.14×10^{-4} mg/cm^2 and 0.02 μmol/cm^2, respectively) of repellent activity predicted by our models. However, all 16 EOCs with 5-6 sensilla activation are substances very similar to compounds with repellent activities previously reported.

It should be mentioned that 31 chemicals from 234 EOCs selected for our models have been previously reported with repellent activity in the literature, which mentioned the majority of the individual chemical components found in EOs used in patented repellent inventions.[87] Most of these compounds showed activation on four sensilla. In fact, only 1 EOC is among 16 compounds with activation of 5/6 sensilla. Among these 31 EOCs previously reported are Camphene, Camphor, Carvone, 1.8-Cineole, Citronellal, Limonene, Linalool, Terpinene-4-ol, Verbenone, and others (for more detail see review[87] and also[55], [64], [89]) which are used as additives in at least one patent and tested against species of mosquito vectors of infectious diseases. Here, it is important to note that the predicted values of RD_{50} for some of these compounds (which had this parameter reported in [87]) are similar to the experimental values reported in the review article.[87] These results are shown in **Table 7**. This

demonstrates that if these oils had not been assessed and we were to select them for evaluation against mosquito species (from chemical composition), they would show repellent activity, which confirms our models' validity.

Table 7. Observed and Predicted RD_{50} Values for EOCs Used in Patented Repellent Inventions that were identified from VS as Potential IRs.

EOCs	**RD_{50} mg/cm^2**	
	Obs.*	**Pred.**
Camphene	0.0022	0.0021
Camphor	0.0014	0.0020
Carvone	0.0013	0.0011
1,8-Cineole	0.014	0.0025
Citronellal	0.00022	0.00083
Limonene	0.0018	0.00060
Linalool	0.0015	0.00058
Terpinen-4-ol	0.0015	0.0020
Verbenone	0.0016	0.0027

*Data collected from.[87]

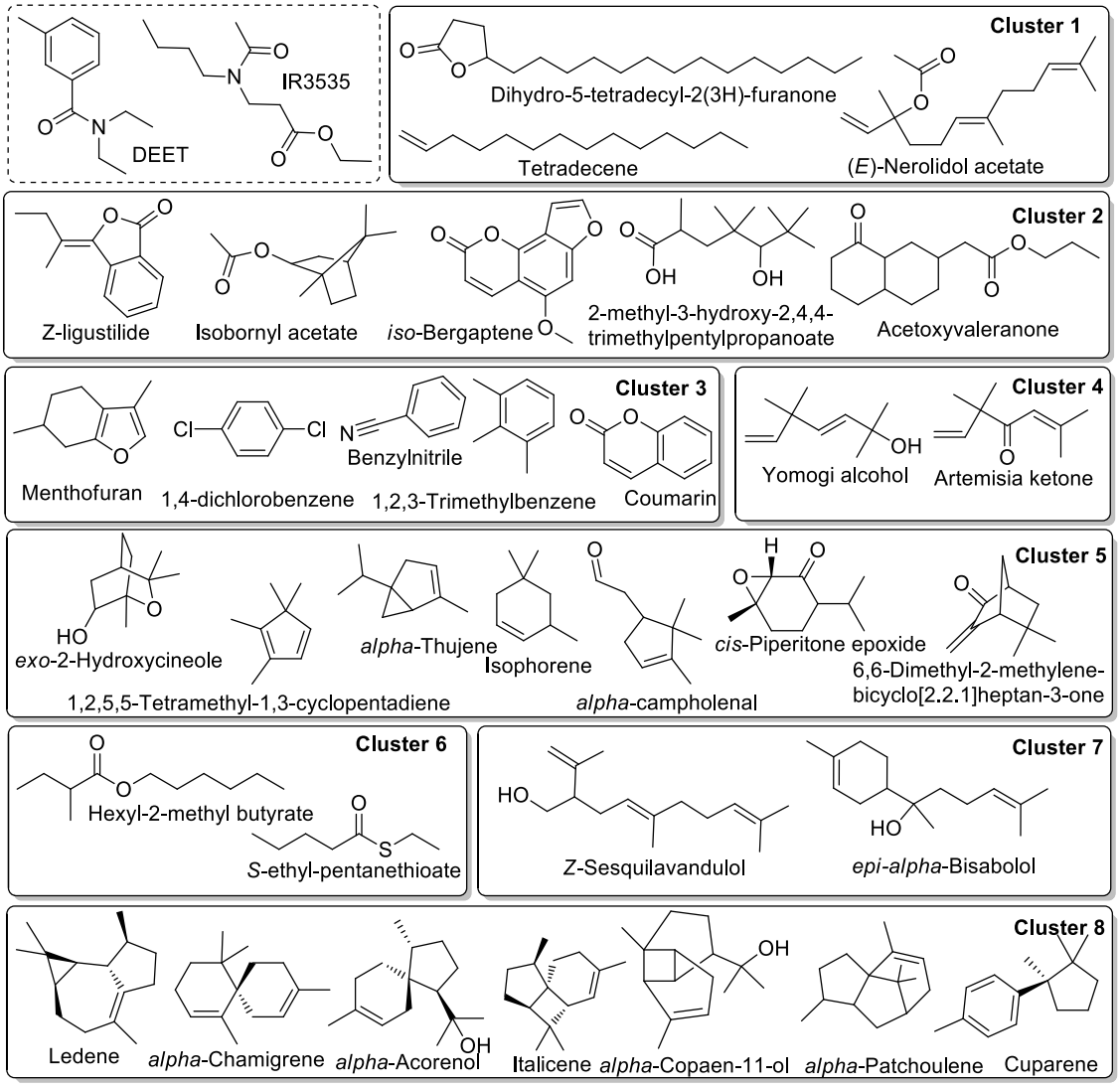
DEET
IR3535
Cluster 1
Dihydro-5-tetradecyl-2(3H)-furanone
Tetradecene
(E)-Nerolidol acetate
Cluster 2
Z-ligustilide
Isobornyl acetate
iso-Bergaptene
2-methyl-3-hydroxy-2,4,4-trimethylpentylpropanoate
Acetoxyvaleranone
Cluster 3
Menthofuran
1,4-dichlorobenzene
Benzylnitrile
1,2,3-Trimethylbenzene
Coumarin
Cluster 4
Yomogi alcohol
Artemisia ketone
Cluster 5
exo-2-Hydroxycineole
1,2,5,5-Tetramethyl-1,3-cyclopentadiene
alpha-Thujene
Isophorene
alpha-campholenal
cis-Piperitone epoxide
6,6-Dimethyl-2-methylene-bicyclo[2.2.1]heptan-3-one
Cluster 6
Hexyl-2-methyl butyrate
S-ethyl-pentanethioate
Cluster 7
Z-Sesquilavandulol
epi-alpha-Bisabolol
Cluster 8
Ledene
alpha-Chamigrene
alpha-Acorenol
Italicene
alpha-Copaen-11-ol
alpha-Patchoulene
Cuparene

Figure 6. Virtual screening and selection of 33 EOCs (virtual hits) with potential repellent activity. The 50 IRs used as internal reference were grouped as follows: Cluster 1) Linoleic acid, Oleic acid, Palmtic acid and Phytol; Cluster 2) Dimethyl phthalate, Dibutyl phthalate, Menthyl acetate, DEET and Permethrin; Cluster 3) Trans-cinnamaldehyde, Cinnamyl alcohol, Eugenol, Thymol, Carvacrol and Naphthalene; Cluster 4) Isoamyl alcohol, Myrcene, Citronellal, Citronellol, (-)-Linalool and Geraniol; Cluster 5) R-(+)-Limonene, S-(-)-Limonene, α-Terpinene, α-Pinene, (+)-α-Pinene, (-)-α-Pinene, (-)-β-Pinene, (+)-β-Pinene, 1S-(+)-3-Carene, Menthoglycol (PMD), S-(-)-Perillaldehyde, S-(-)-perillyl alcohol, (-)-Menthone, (+)-Menthone, α-Terpineol, (+)-Terpinen-4-ol, D-neomenthol, Menthol, Terpinolene, S-cis-Verbenol, Camphor and Eucalyptol; Cluster 6) Citronellic acid, Geranyl acetone and Linalyl acetate; Cluster 7) Geranyl acetate; and Cluster 8) β-Caryophyllene and (-)-Caryophyllene oxide. The structures of the IRs DEET and IR3535 are also presented.

Taken representatively from 8 clusters and considering the structural di-similarity with well-known IRs, we have selected 33 compounds that could exhibit significant repellent activity and fulfill the important goals of an ideal IR, having a relatively novel structure core (see **Figure 6**). Compounds in **Figure 6** are an example of the main frameworks with theoretical repellent activities, and some of them are new possible hits; the majority of the EOCs in **Figure 6** produced a theoretical response in at least two sensilla, which is an adequate result if taken into account that well known IRs show significant activation of only one sensillum (e. g., DEET on ORNs in SST-C).[29]

Finally, we developed an analysis of the leading frameworks in each cluster and looked in-depth at several literature sources to draw the main conclusions about structural features in the 33 EOCs in **Figure 6** and verify whether they represent new chemotypes.

Cluster 1 included Linoleic acid, Oleic acid, Palmitic acid, and Phytol plus 23 EOCs from 234 compounds selected in the VS experiment (**EXP3**). Here, only one EOC (Tetradecene) showed theoretical activation on five different types of sensilla. This compound has not been reported in any insect repellent activity, but two position isomers, 6- and 7-tetradecane, have been recognized as alarm pheromones, which is, in fact, a "repellent effect".[132] All compounds that showed actions against 3 or 4 different types of sensilla are structurally very close to some IR frameworks. Amongst the EOCs with strong action on 2 sensilla are Dihydro-5-tetradecyl-2(3H)-furanone (antitubercular activity against *Mycobacterium tuberculosis*)[133] and 4,8,12,16-tetramethylheptadecan-4-olide. These compounds have a new chemotype (isoprenoid γ-lactone) because no reports exist with their repellent activities and are structurally rather different to well-known IRs or related substances with action against insects. Finally, in this cluster, several monoterpenoids exist (C-10, two isoprene units), like (*E*)-Nerolidol, which is an acetate derivative of the IR Fokienol.

In the **second cluster,** the IRs: Dimethyl phthalate, Dibutyl phthalate, Menthyl acetate, DEET, and Permethrin were comprised in join with 13 EOCs (1-4 sensilla). Here exist several acetate-derived chemicals related to the IRs Menthyl acetate (like Isobornyl acetate (directly related to with the IR Borneol),[87] cis-carvyl acetate, isodihydrocarveol acetate, Isobornyl acetate, etc.). In fact, isobornyl acetate possesses repellent activity, confirmed against mosquitoes and moths.[87] On the other hand, compounds like (3Z)-butylidene phthalide (1 sensillum) and *Z*-ligustilide (2 sensilla, see **Figure 6**) are a representative framework in this cluster. Z-ligustilide (LIG, (3Z)-3-butylidene-4,5-dihydroisobenzofuran-1-one), an essential oil extract from *Angelica Sinensis*, has broad pharmaceutical applications in treating cardiovascular diseases and ischemic brain injury. LIG can inhibit the proliferation and cell cycle progression of vascular smooth muscle

cells, associated with basic fibroblast growth factor stimulation, through the reduction of reactive oxygen species and/or the suppression of the Mitogen-Activated Protein Kinases pathway (MAPK). Meanwhile, LIG has an analgesic effect on rats and a concentration-dependent anti-inflammatory effect on lipopolysaccharide-activated rat microglia without cytotoxicity.[134] Since the IR activity of LIG was not reported previously, the results of this study should raise public concern over the potential application associated with insect behavior.

Other relevant scaffolds in **Cluster 2** are represented by *iso*-Bergaptene and Acetoxyvaleranone, for which the repellent action against any insect has not been reported yet. However, the first EOC is a furocoumarin that presents photo-toxicity and, therefore, is still advisable to keep treated skin out of the sun and to use it in concentrations of less than 1%.[135] Thus, only Acetoxyvaleranone will be declared as a representation of a new chemotype.

Cluster 3 is characterized by *trans*-cinnamaldehyde, Cinnamyl Alcohol, Eugenol, Thymol, Carvacrol, and Naphthalene as IRs and included 74 EOCs. This cluster is determined by molecules with a cyclic aromatic system, like 1,2,3-Trimethylbenzene (and its isomers), Benzylnitrile, Phenol, Menthofuran, Coumarin, and so on. Only 1,3,5-trimethyl benzene produces a response in 5 different types of sensilla. However, this compound and some isomers are central nervous system depressants and may cause respiratory disorders. The 1,2,4-isomer may also be narcotic. Other effects of exposure to these compounds include headaches, tension, nervousness, inflammation and hemorrhaging of mucous membranes, convulsions, and ultimately death.[136] The major chemicals in this cluster are structurally very close to IRs, for example, 4-vinyl-*o*-guaiacol. The most diverse compounds here: Dichlorobenzene and Menthofuran, possess repellent activity confirmed against mosquitoes and moths.[87], [137], [138] In

fact, 1,4-dichlorobenzene is formulated for moth repellents but presents some problems related to human and environmental health such as; irritating to eyes, it is highly toxic and may cause long-term adverse effects in aquatic environments, and has carcinogenic effects.[139]

The IRs: Isoamyl alcohol, Myrcene, Citronellal, Citronellol, (-)-Linalool, and Geraniol were grouped in **Cluster 4**. All EOCs in this cluster are also alcohols or carbonyls with aliphatic skeletons, like *cis*-Salvene, *cis*-Ocimene, etc., very close to the IR Myrcene. However, two related compounds exist, Artemisia ketone and Yomogi alcohol (see **Figure 6**), which will be considered here as a new terpene-based chemotype. In fact, some studies of the chemical composition of plant extracts (oil from *Artemisia argyi*, *A. feddei*, *A. gmelinii*, *Tanacetum vulgare*)[140]–[142] with biological activity against arthropods (insects, but also ticks) and patented inventions report the presence of these EOCs. However, none of the compounds have been evaluated individually in these experiments. One of the recent studies show that Artemisia ketone is the oil component that has the most excellent anti-microbial activity; in fact, it always turns out to be effective against bacteria and some fungi (*Candida albicans* and *Aspergillus fumigatus*) at low concentrations (range 0.07–10.0mg/mL).[142] Another study evidenced that the presence of Artemisia ketone in the blend caused a significant increase in the repellency of the resulting blend (essential oil of *Suregada zanzibariensis* leaves) and that some blends of terpenoid ketones can serve as effective *A. gambiae* mosquito repellents.[143]

Cluster 5 have the following IRs: *R*-(+)-Limonene, *S*-(-)-Limonene, α-Terpinene, α-Pinene, (+)-α-Pinene, (-)-α-Pinene, (-)-β-Pinene, (+)-β-Pinene, 1*S*-(+)-3-Carene, Menthoglycol (PMD), *S*-(-)-Perillaldehyde, *S*-(-)-Perillyl alcohol, (-)-Menthone, (+)-Menthone, α-Terpineol, (+)-Terpinen-4-ol, *D*-neomenthol, Menthol, Terpinolene, *S-cis*-Verbenol, Camphor, and Eucalyptol. In addition, 52 EOCs were also included. This

cluster has the compound with most responses against mosquito antennal sensilla, where 4 and 7 EOCs exist with action on 6 and 5 different types of sensilla (Camphene, α-Thujene, *E*-Pinene hydrate, and Thuja-2,4-10-diene) and 7 (*cis*-Sabinene hydrate, α-Isophorone, 5-methylenenorbonene, *cis*-1,4-Dimethylcyclohexane, 1,2,5,5-Tetramethyl-1,3-cyclopentadiene, *trans*-2-Methyldecalin, Isophorone), respectively. The compounds with maximal responses are very close to well-known IRs, where also are Sabinene, *neo*-3-thujanol, thuja-2,4-10-diene, α-thujenal, thuj-3-en-10-al, etc. However, that structural framework is not new in the IR field. A similar case occurs with the molecule 6,6-Dimethyl-2-methylene-bicyclo[2.2.1] heptane-3-one (5 sensilla) due to them being very close to Pinene isomers *S*-*cis*-Verbenol, Camphor, etc. Menthone isomers are other IR chemotypes in this cluster, and *cis*-Piperitone epoxide (4 sensilla) represents this kind of structure. Another kind of structure is represented by Isophorene that produces responses in 5 sensilla too, but is not a new scaffold because this type of skeleton has a good representation in IR panel, for instance, Limonene isomers, 3-Carene isomers, etc. (see **Figure 6**). Here, only the 5-C ring molecules 1,2,5,5-Tetramethyl-1,3-cyclopentadiene (5 sensilla; also 5-*tert*-Butyl-1,3-cyclopentadiene (2 sensilla)) and α-campholenal (3 sensilla) have a new chemotype (see **Figure 6**).

In **Cluster 6**, only 13 EOCs are grouped with Citronellic acid, Geranyl acetone, and Linalyl acetate. Here only the molecule of *S*-ethyl-pentanethioate has an interesting structure (dissimilar) regarding well-known IRs, which is not very similar to merck790 and 2-undecanone. However, the Hexyl-2-methyl butyrate is a skin irritant, and its decomposition with heat is dangerous for the environment. Thus, *S*-ethyl-pentanethioate can be considered as a new theoretical IR structural prototype.

In **Cluster 7,** only one IR of reference is included, Geranyl acetate (Z-Sesquilavandulol is a structurally similar EOC), which is joint with 23 EOCs.

Compounds of the bicyclic sesquiterpene class, such as 7-*epi*-Sesquithujene were selected as a good prototype, and in fact, this EOC has been shown to elicit solid electrophysiological responses on the antennae of the emerald ash borer, *Agrilus planipennis*.[144] Another much related EOC was *epi*-α-bisabolol (monocyclic sesquiterpene alcohol with good theoretical results), which has a slight sweet flower fragance. Also its use since ancient times is due to the healing properties on the skin, which is why it is used in the cosmetic industry. Furthermore, it is known to have anti-irritant, anti-inflammatory, and anti-microbial properties.[145] It also indicates improvement in UV-induced skin damage and promotes skin growth. Scientific studies have also shown that α-Bisabolol has excellent anti-obesity and anti-oxidation effects. In addition, this compound showed good insect repellent activity (84.0%)[87], [103], and due to its good properties, it is considered a good IR lead (starting point). In fact, other EOCs in their chemical neighborhood like Turmerona showed 88.9% of spatial repellency for 180 min (female *A. aegypti*).[87], [103]A Turmerona-related compound selected by us as a good IR candidate is β-Atlantone (2 sensilla and good repellent activity predicted).

Finally, **Cluster 8** had two of the most potent plant-related IRs, β-Caryophyllene and (-)-Caryophyllene oxide (both have responses in 4 different sensilla types). In this cluster, several bicyclic systems exist, like Cuparene, Humulene, α-Patchoulene (3 sensilla; and Copaeb-11-ol, four sensilla), spiroalcanes (Italicene (3 sensilla), α-Carenol (2 sencilla) and Chamigrene (3 sensilla)), Ledene (5 sensilla) and santalos, which are viable scaffolds for developing more diverse active IRs. However, Santalos (1 sensillum) is an EOC with many reports of insect repellency[45]. In that cluster, there are two analogs with better theoretical activity (*Cis*-Sesquisalebinene; 4 sensilla) and (Campherene; 2 sensilla). Ledene is structurally similar to the IR Spathulenol. However,

Cuparene, which is also recognized as an IR, is structurally similar to the skeleton of 2-Phenylcyclohexanol (X_4).

On the other hand, numerous studies of chemical composition of plant extracts with biological activity against arthropods (insects, but also ticks) and patented inventions, report the presence of most of the compounds mentioned in our list, such is the case of extracts from *Pulicaria gnaphalodes* and *Achillea wilhelmsii* (dehydro aromadendrene, chrysanthenyl acetate),[146] extract from the turmeric herb (α-Copaen-11-ol),[147] oil from *Artemisia argyi*, *A. feddei*, *A. gmelinii*, *Tanacetum vulgare* (*cis*-Carvyl acetate, *cis*-Chrysanthenyl acetate, Artemisia ketone, Yomogi alcohol, Zingiberenol, 6,6-Dimethyl-2-methylene-bicyclo[2.2.1] heptan-3-one),[140], [141] oil from the root of *A. annua* (*E*-Nerolidol acetate, (*E*,*E*)-Farnesyl acetate),[148] oil from *Tagetes minuta* (*cis*-Ocimene),[149] extracts of *Plectranthus incanus* (*cis*-Piperitone epoxide),[150] oil from *Cedrus deodara* (β-Atlantone),[151] extract from *Guava* fruit (Isoamyl 2-methylbutyrate),[152] *Angelica sinensis* oil ((*Z*)-3-butylidene phthalide),[153] oil from *Croton roxburghii* (α-Campholenal),[154] and oils from Mediterranean plants (Hexyl 2-methyl butyrate).[155] Although none of the compounds mentioned above have been evaluated individually, this suggests that the extracts' repellent activity may be associated with the presence of these metabolites. However, compounds such as Tetradecene, Dihydro-5-tetradecyl-2(3H)-furanone, (3Z)-butylidene phthalide, Acetoxyvaleranone, Yomogi alcohol, 1,2,5,5-Tetramethyl-1,3-cyclopentadiene, α-campholenal, *S*-ethyl-pentanethioate, β-Atlantone, Humulene, Copaeb-11-ol, Italicene, α-Carenol, Chamigrene, and Ledene their repellent activity is reported here for the first time. It, therefore, is a successful expansion of the existing scaffold. These EOCs are unknown for the developed models, so we can say that their evaluation in the equations is equivalent to discovering new IR leads. Furthermore,

according to the literature sources consulted, none of them have been yet described as IRs. These suggestive results need to be confirmed by experimental tests to obtain the final and conclusive results about OBPs/ORs/ORNs and/or insect repellent profiles of these compounds.

As the second part of **EXP3**, it was used the Malaria Box [66] library (available in Supporting Information folder **SI_4**) for VS-prospective analysis using SiliS-PAPACS software. Malaria Box comprises a group of 400 drug-like and probe-like compounds found in an exhaustive search to combat malaria and neglected diseases in 2011[66]. From this starting point of 400 compounds, only 80 of them were filtered based on the molecular weight (MW) and the low volatility it implies, being an undesired feature in IRs design (see **SI_4_EXP3_ResultsMBox_filtering_MW** file). After docking the 80 molecules by using SiliS-PAPACS software with the parameters as default, there were filtered to keep only the active molecules and the ones with adequate volatility (see **SI_4_EXP3_ResultsMBox_VS** file). Finally, eight molecules were found to be ideal to use in IRs search; these are presented in **Figure 7**. In these cluster, half are probe-like compounds: MMV000570 (EC_{50} = 194nM, MW = 278.35), MMV665812 (EC_{50} = ND, MW = 255.35), MMV665924 (EC_{50} = 1060nM, MW = 286.75), & MMV666021 (EC_{50} = 94.4nM, MW = 272.30). And the other half are drug-like compounds: MMV000911 (EC_{50} = ND, MW = 278.71), MMV008270 (EC_{50} = ND, MW = 265.31), MMV018984 (EC_{50} = 693nM, MW = 278.31), & MMV665820 (EC_{50} = 520nM, MW = 293.53).

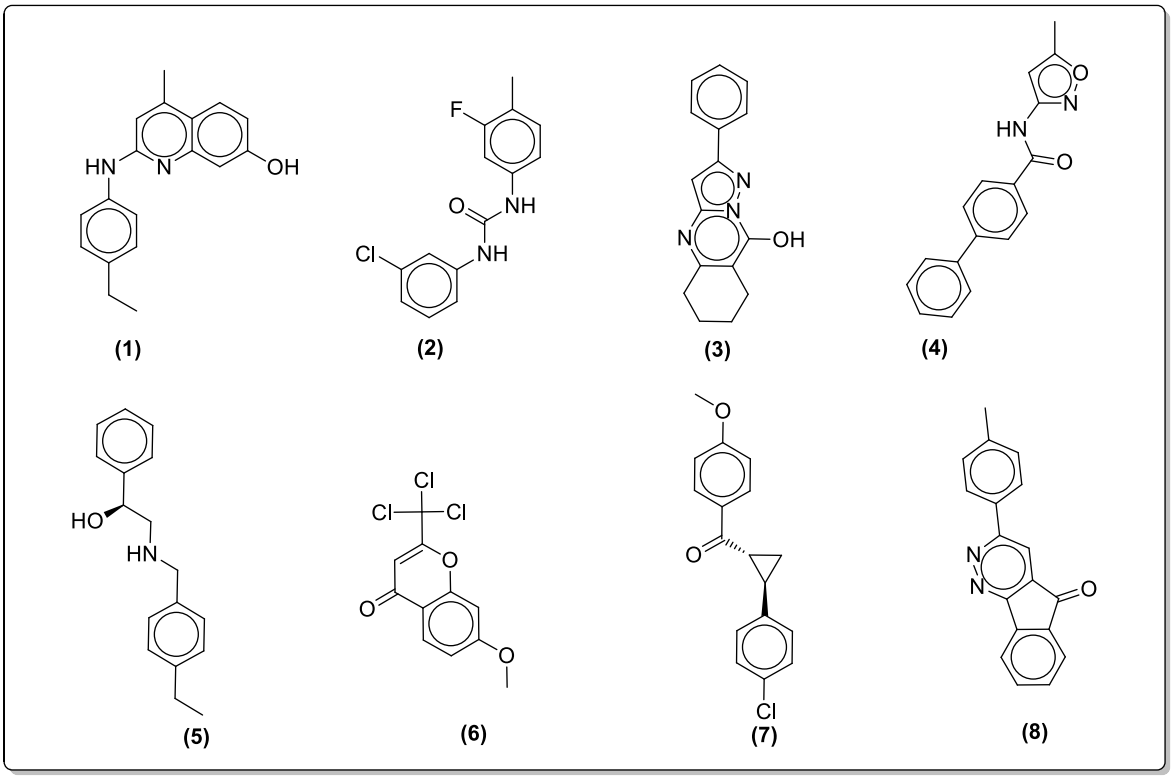

Figure 7. Molecules selected from Malaria Box[66] library through VS using SiliS-PAPACS software. 1: MMV000570, 2: MMV000911, 3: MMV008270, 4: MMV018984, 5: MMV665812, 6: MMV665820, 7: MMV665924, 8: MMV666021.

As shown in **Figure 7**, these compounds present different structures from those studied to date in the search for repellents, including DEET. Besides, until now, no reports of application have been found to search for repellents of any of these eight compounds.

4. CONCLUDING REMARKS

This report's most outstanding contribution is that 23 compounds have been found starting from libraries of multiple compounds, among them compounds of essential oils and the well-known Malaria box. These 23 compounds are presented as new and key scaffolds for the research and development of the ideal repellent against

different vector-borne insects; however, they are suggested to be subjected to experimentation in the laboratory and *in situ*.

In this report, we developed an expert system named SiliS-PAPACS that internally implements the workflow of different programs that, employing QSAR models based on MDs, and structure models based on AV docking affinities, allow to obtain quantitative models that discriminate and predict the ORN responses in six different types of the antennal trichoid sensilla of *C. quinquefasciatus* to estimate IR activities. The developed models were used in VS to simulate from the *in silico* to 'real-world applications. This is presented as a novelty that optimizes time and costs by saving experimentation steps and reproducing reliable computationally-assisted results. Besides, the *biosilico* identification of novel potential EOCs as IRs is reported by using the new computational pipeline, and action/activity profiles appear to be promising from a practical point of view. We propose that these models are adequately characterized and evaluated and could play a valuable and essential role in the early discovery of new compounds with potent repellent activity; that is to say, the present study has laid a foundation for developing olfactory-based insect control agents.

5. FUTURE OUTLOOKS

Regarding the promising behavior that this computational approach brings through SiliS-PAPACS, we seek to study more diverse libraries of compounds to find leading ones with ideal properties. Furthermore, instead of modeling the sensilla response, we will analyze the ML techniques used in this report in experimental results of compounds with repellent properties against different insect species' odorant receptors.

SUPPORTING INFORMATION

The Supporting Material of the present report is available in the following Google Drive folder:

https://drive.google.com/drive/folders/1BMOaYgNgHqYwWiEuYQHZ9Ij3BuTSasHO?usp=sharing

6. REFERENCES

[1] G. G. E. Scudder, *Insect Biodivers. Sci. Soc.*, **2009**, DOI:10.1002/9781444308211.ch2.

[2] A. F. Carey, J. R. Carlson, *Proc. Natl. Acad. Sci. U. S. A.*, **2011**, DOI:10.1073/pnas.1103472108.

[3] J. C. Dickens, J. D. Bohbot, *Insect Repellents Handbook, Second Ed.*, **2014**, DOI:10.1201/b17407.

[4] M. Brown, A. A. Hebert, *J. Am. Acad. Dermatol.*, **1997**, DOI:10.1016/S0190-9622(97)70289-5.

[5] United States Environmental Protection Agency, DEET, https://www.epa.gov/insect-repellents/deet#:~:text=It is widely used to,and Rocky Mountain spotted fever.

[6] S. Majeed, S. R. Hill, R. Ignell, *J. Exp. Biol.*, **2014**, DOI:10.1242/jeb.092718.

[7] J. S. Ignatious Raja, N. Katanayeva, V. L. Katanaev, C. G. Galizia, *Eur. J. Neurosci.*, **2014**, DOI:10.1111/ejn.12481.

[8] P. Xu, F. Zhu, G. K. Buss, W. S. Leal, *F1000Research*, **2015**, DOI:10.12688/f1000research.6646.1.

[9] I. Vythilingam, G. L. Chiang, S. T. Chan, *Southeast Asian J. Trop. Med. Public Health*, **1992**, *23*, 328–331.

[10] D. L. Kline, *J. Am. Mosq. Control Assoc.*, **1994**, *10*, 280–287.

[11] J. I. Raji, N. Melo, J. S. Castillo, S. Gonzalez, V. Saldana, M. C. Stensmyr, M. DeGennaro, *Curr. Biol.*, **2019**, DOI:10.1016/j.cub.2019.02.045.

[12] M. Geier, O. Bosch, B. Steib, A. Rose, *Proc. 4th Int. Conf. Urban Pests.*, **2002**, 37–46.

[13] S. A. Allan, U. R. Bernier, D. L. Kline, *J. Vector Ecol.*, **2006**,

DOI:10.3376/1081-1710(2006)31[71:aomtva]2.0.co;2.

[14] S. S. Sucharita, In silico screening and molecular docking of phytochemical compounds to identify novel mosquito/insect repellent compounds targeting the odorant binding proteins (OBPs) of Anopheles gambiae and Anopheles stephensi, Orissa University of Agriculture and Technology, 2016.

[15] H. Venthur, J. Zhou, **2018**, DOI:10.3389/fphys.2018.01163.

[16] Z. Hubálek, *Emerg. Infect. Dis.*, **2003**, *9*.

[17] K. R. Chauhan, U. R. Bernier, *Insect Repellents Handbook, Second Ed.*, **2014**, DOI:10.1201/b17407.

[18] J. Wang, F. Zhu, X. M. Zhou, C. Y. Niu, C. L. Lei, *J. Stored Prod. Res.*, **2006**, DOI:10.1016/j.jspr.2005.06.001.

[19] P. Xu, Y. M. Choo, A. De La Rosa, W. S. Leal, *Proc. Natl. Acad. Sci. U. S. A.*, **2014**, DOI:10.1073/pnas.1417244111.

[20] B. W. Bissinger, R. M. Roe, *Pestic. Biochem. Physiol.*, **2010**, DOI:10.1016/j.pestbp.2009.09.010.

[21] G. Briassoulis, *Hum. Exp. Toxicol.*, **2001**, DOI:10.1191/096032701676731093.

[22] L. C. Rutledge, M. A. Moussa, C. A. Lowe, R. K. Sofield, *J. Med. Entomol.*, **1978**, DOI:10.1093/jmedent/14.5.536.

[23] E. J. Murphy, J. C. Booth, F. Davrazou, A. M. Port, D. N. M. Jones, *J. Biol. Chem.*, **2013**, DOI:10.1074/jbc.M112.436386.

[24] N. P. Charlton, L. T. Murphy, J. L. Parker Cote, J. P. Vakkalanka, *Clin. Toxicol.*, **2016**, DOI:10.1080/15563650.2016.1186806.

[25] O. Anderbrant, Pheromones, *Encyclopedia of Ecology*. Elsevier, , 2707–2709, 2008.

[26] M. Tegoni, V. Campanacci, C. Cambillau, *Trends Biochem. Sci.*, **2004**,

DOI:10.1016/j.tibs.2004.03.003.

[27] N. M. Abd El-Ghany, Pheromones and Chemical Communication in Insects, *Pests, Weeds and Diseases in Agricultural Crop and Animal Husbandry Production*. IntechOpen, , 2020.

[28] H. Song, J. Y. Kwon, H. S. Han, Y. Bae, C. Moon, **2008**, DOI:10.3390/s8106303.

[29] F. Liu, L. Chen, A. G. Appel, N. Liu, *J. Insect Physiol.*, **2013**, DOI:10.1016/j.jinsphys.2013.08.016.

[30] J. Zhou, DOI:10.1016/S0083-6729(10)83010-9.

[31] M. Maïbèche-Coisne, F. Sobrio, T. Delaunay, M. Lettere, J. Dubroca, E. Jacquin-Joly, P. Nagnan-Le Meillour, *Insect Biochem. Mol. Biol.*, **1997**, DOI:10.1016/S0965-1748(96)00088-4.

[32] G. Du, G. D. Prestwich, *Biochemistry*, **1995**, DOI:10.1021/bi00027a023.

[33] M. S. Kim, A. Repp, D. P. Smith, *Genetics*, **1998**, *150*, 711–721.

[34] R. Dhivya, K. Manimegalai, *Int. J. Pharm. Phytopharm. Res.*, **2013**, *3 (2)*, 134–138.

[35] R. G. Vogt, F. E. Callahan, M. E. Rogers, J. C. Dickens, *Chem. Senses*, **1999**, DOI:10.1093/chemse/24.5.481.

[36] P. Pelosi, M. Cereda, G. Foti, M. Giacomini, M. Pesenti, *Am. J. Respir. Crit. Care Med.*, **1995**, DOI:10.1164/ajrccm.152.2.7633703.

[37] V. Sargsyan, M. N. Getahun, S. L. Llanos, S. B. Olsson, B. S. Hansson, D. Wicher, *Front. Cell. Neurosci.*, **2011**, DOI:10.3389/fncel.2011.00005.

[38] R. Farbiszewski, R. Krancc, *Polish Ann. Med.*, **2013**, DOI:10.1016/j.poamed.2013.02.002.

[39] Zainulabeuddin Syed, *Insect Repellents Handbook, Second Ed.*, **2014**, 43–52.

[40] M. E. Rogers, J. Krieger, R. G. Vogt, *J. Neurobiol.*, **2001**, DOI:10.1002/neu.1065.

[41] C. E. Merrill, J. Riesgo-Escovar, R. J. Pitts, F. C. Kafatos, J. R. Carlson, L. J. Zwiebel, *Proc. Natl. Acad. Sci. U. S. A.*, **2002**, DOI:10.1073/pnas.022505499.

[42] I. Ishida, W. S. Leal, *Insect Biochem. Mol. Biol.*, **2002**, DOI:10.1016/S0965-1748(02)00136-4.

[43] K. Sato, M. Pellegrino, T. Nakagawa, T. Nakagawa, L. B. Vosshall, K. Touhara, *Nature*, **2008**, DOI:10.1038/nature06850.

[44] P. J. Clyne, C. G. Warr, M. R. Freeman, D. Lessing, J. Kim, J. R. Carlson, *Neuron*, **1999**, DOI:10.1016/S0896-6273(00)81093-4.

[45] Z. Syed, W. S. Leal, *Proc. Natl. Acad. Sci. U. S. A.*, **2009**, DOI:10.1073/pnas.0906932106.

[46] D. Restrepo, J. H. Teeter, D. Schild, *J. Neurobiol.*, **1996**, DOI:10.1002/(SICI)1097-4695(199605)30:1<37::AID-NEU4>3.0.CO;2-H.

[47] M. Schmuker, G. Schneider, *Proc. Natl. Acad. Sci. U. S. A.*, **2007**, DOI:10.1073/pnas.0705683104.

[48] K. S. da Costa, J. M. Galúcio, C. H. Souza da Costa, A. R. Santana, V. dos Santos Carvalho, L. D. do Nascimento, A. H. Lima e Lima, J. N. Cruz, A. Lameira, C. N. Alves, J. Lameira, **2019**, DOI:10.1021/acsomega.9b03157.

[49] J. S. Portilla-Pulido, R. M. Castillo-Morales, M. A. Barón-Rodríguez, J. E. Duque, S. C. Mendez-Sanchez, *J. Med. Entomol.*, **2020**, DOI:10.1093/jme/tjz171.

[50] J. V. Gopal, K. Krishnan, **2013**, DOI:10.1007/s12539-013-0152-2.

[51] M. I. Qadir, M. Arshad, **2015**.

[52] P. D. K. Jayanthi, V. Kempraj, R. Aurade, *Pest Manag. Hortic. Ecosyst.*, **2016**, *22*, 20–27.

[53] T. Thireou, G. Kythreoti, K. E. Tsitsanou, K. Koussis, C. E. Drakou, J. Kinnersley, T. Kröber, P. M. Guerin, J. J. Zhou, K. Iatrou, E. Eliopoulos, S. E. Zographos, *Insect Biochem. Mol. Biol.*, **2018**, DOI:10.1016/j.ibmb.2018.05.001.

[54] A. R. Katritzky, Z. Wang, S. Slavov, M. Tsikolia, D. Dobchev, N. G. Akhmedov, C. D. Hall, U. R. Bernier, G. G. Clark, K. J. Linthicum, *Proc. Natl. Acad. Sci. U. S. A.*, **2008**, DOI:10.1073/pnas.0800571105.

[55] G. Paluch, J. Grodnitzky, L. Bartholomay, J. Coats, *J. Agric. Food Chem.*, **2009**, DOI:10.1021/jf900964e.

[56] P. V. Oliferenko, A. A. Oliferenko, G. I. Poda, D. I. Osolodkin, G. G. Pillai, U. R. Bernier, M. Tsikolia, N. M. Agramonte, G. G. Clark, K. J. Linthicum, A. R. Katritzky, G. Girinath, U. R. Bernier, M. Tsikolia, N. M. Agramonte, G. G. Clark, K. J. Linthicum, A. R. Katritzky, *PLoS One*, **2013**, DOI:10.1371/journal.pone.0064547.

[57] G. Caballero-Vidal, C. Bouysset, H. Grunig, S. Fiorucci, N. Montagné, J. Golebiowski, E. Jacquin-Joly, *Sci. Rep.*, **2020**, DOI:10.1038/s41598-020-58564-9.

[58] J. I. B. Janairo, G. C. Janairo, F. F. Co, *Nov. Biotechnol. Chim.*, **2018**, DOI:10.2478/nbec-2018-0006.

[59] D. Kepchia, P. Xu, R. Terryn, A. Castro, S. C. Schürer, W. S. Leal, C. W. Luetje, *Sci. Rep.*, **2019**, DOI:10.1038/s41598-019-40640-4.

[60] J. R. Valdés-Martiní, Y. Marrero-Ponce, C. R. García-Jacas, K. Martinez-Mayorga, S. J. Barigye, Y. S. Vaz D'Almeida, H. Pham-The, F. Pérez-Giménez, C. A. Morell, *J. Cheminform.*, **2017**, DOI:10.1186/s13321-017-0211-5.

[61] C. R. García-Jacas, Y. Marrero-Ponce, L. Acevedo-Martínez, S. J. Barigye, J. R. Valdés-Martiní, E. Contreras-Torres, *J. Comput. Chem.*, **2014**,

DOI:10.1002/jcc.23640.

[62] O. Trott, A. J. Olson, *J. Comput. Chem.*, **2010**, DOI:10.1002/jcc.21334.AutoDock.

[63] J. M. Gaudin, T. Lander, O. Nikolaenko, *Chem. Biodivers.*, **2008**, DOI:10.1002/cbdv.200890058.

[64] M. O. Omolo, D. Okinyo, I. O. Ndiege, A. Hassanali, **2004**, DOI:10.1016/j.phytochem.2004.08.035.

[65] T. N. C. Wells, *Malar. J.*, **2011**, DOI:10.1186/1475-2875-10-S1-S3.

[66] T. Spangenberg, J. N. Burrows, P. Kowalczyk, S. McDonald, T. N. C. Wells, P. Willis, *PLoS One*, **2013**, DOI:10.1371/journal.pone.0062906.

[67] RDKit: Open-source cheminformatics. .

[68] R. Medina Marrero, Y. Marrero-Ponce, S. J. Barigye, Y. Echeverría Díaz, R. Acevedo-Barrios, G. M. Casañola-Martín, M. García Bernal, F. Torrens, F. Pérez-Giménez, *SAR QSAR Environ. Res.*, **2015**, DOI:10.1080/1062936X.2015.1104517.

[69] J. R. Mora, Y. Marrero-Ponce, C. R. García-Jacas, A. Suarez Causado, *Chem. Res. Toxicol.*, **2020**, DOI:10.1021/acs.chemrestox.0c00030.

[70] R. W. P. Urias, S. J. Barigye, Y. Marrero-Ponce, C. R. García-Jacas, J. R. Valdes-Martiní, F. Perez-Gimenez, *Mol. Divers.*, **2015**, DOI:10.1007/s11030-014-9565-z.

[71] G. Wolber, T. Langer, *J. Chem. Inf. Model.*, **2005**, DOI:10.1021/ci049885e.

[72] L. Schrödinger, The PyMOL Molecular Graphics System. 2002.

[73] E. Frank, M. A. Hall, I. H. Witten, in The WEKA Workbench, **2016**.

[74] R. Kohavi, G. H. John, *Artif. Intell.*, **1997**, DOI:10.1016/S0004-3702(97)00043-X.

[75] R. Hernández-Lambraño, Y. Marrero-Ponce, F. Peréz-Giménez, Y. Perez-Castillo, The Role of the Odorant Binding Proteins for Insect Repellents: Docking Studies of Agents that Interfere with Olfaction. 2021.

[76] P. Baldi, S. Brunak, Y. Chauvin, C. A. F. Andersen, H. Nielsen, *Bioinformatics*, **2000**, DOI:10.1093/bioinformatics/16.5.412.

[77] M. Sokolova, G. Lapalme, *Inf. Process. Manag.*, **2009**, DOI:10.1016/j.ipm.2009.03.002.

[78] N. V. Chawla, K. W. Bowyer, L. O. Hall, W. P. Kegelmeyer, *J. Artif. Intell. Res.*, **2002**, *16*, 321–357.

[79] L. I. Kuncheva, in Combining Pattern Classifiers: Methods and Algorithms, **2014**.

[80] C. Valsecchi, F. Grisoni, V. Consonni, D. Ballabio, *J. Chem. Inf. Model.*, **2020**, DOI:10.1021/acs.jcim.9b01057.

[81] H. Dragos, M. Gilles, V. Alexandre, *J. Chem. Inf. Model.*, **2009**, DOI:10.1021/ci9000579.

[82] F. Sahigara, K. Mansouri, D. Ballabio, A. Mauri, V. Consonni, R. Todeschini, *Molecules*, **2012**, DOI:10.3390/molecules17054791.

[83] K. Roy, S. Kar, P. Ambure, *Chemom. Intell. Lab. Syst.*, **2015**, DOI:10.1016/j.chemolab.2015.04.013.

[84] J. Jaworska, N. Nikolova-Jeliazkova, *SAR QSAR Environ. Res.*, **2007**, DOI:10.1080/10629360701306050.

[85] Jmol: an open-source Java viewer for chemical structures in 3D, http://www.jmol.org/.

[86] A. R. Katritzky, Z. Wang, S. Slavov, D. A. Dobchev, C. D. Hall, M. Tsikolia, U. R. Bernier, N. M. Elejalde, G. G. Clark, K. J. Linthicum, *J. Med. Entomol.*, **2010**,

DOI:10.1603/ME09284.

[87] A. M. Pohlit, N. P. Lopes, R. A. Gama, W. P. Tadei, V. de Andrade Neto, V. Ferreira, D. A. Neto, *Planta Med.*, **2011**, DOI:10.1055/s-0030-1270723.

[88] T. M. Katz, J. H. Miller, A. A. Hebert, *J. Am. Acad. Dermatol.*, **2008**, DOI:10.1016/j.jaad.2007.10.005.

[89] J. O. Odalo, M. O. Omolo, H. Malebo, J. Angira, P. M. Njeru, I. O. Ndiege, A. Hassanali, *Acta Trop.*, **2005**, DOI:10.1016/j.actatropica.2005.06.007.

[90] Z. Wang, J. Song, J. Chen, Z. Song, S. Shang, Z. Jiang, Z. Han, *Bioorganic Med. Chem. Lett.*, **2008**, DOI:10.1016/j.bmcl.2008.03.091.

[91] P. Pelosi, J. J. Zhou, L. P. Ban, M. Calvello, *Cell. Mol. Life Sci.*, **2006**, DOI:10.1007/s00018-005-5607-0.

[92] S. J. Liu, N. Y. Liu, P. He, Z. Q. Li, S. L. Dong, L. F. Mu, *Arch. Insect Biochem. Physiol.*, **2012**, DOI:10.1002/arch.21036.

[93] K. E. Tsitsanou, T. Thireou, C. E. Drakou, K. Koussis, M. V. Keramioti, D. D. Leonidas, E. Eliopoulos, K. Iatrou, S. E. Zographos, *Cell. Mol. Life Sci.*, **2012**, DOI:10.1007/s00018-011-0745-z.

[94] StatSoft, STATISTICA (data analysis software system). Tulsa, 2001.

[95] Y. Kuwabara, G. V. Alexeeff, R. Broadwin, A. G. Salmon, *Environ. Health Perspect.*, **2007**, DOI:10.1289/ehp.9848.

[96] S. Li, J. F. Picimbon, S. Ji, Y. Kan, Q. Chuanling, J. J. Zhou, P. Pelosi, *Biochem. Biophys. Res. Commun.*, **2008**, DOI:10.1016/j.bbrc.2008.05.064.

[97] A. C. Anderson, *Chem. Biol.*, **2003**, DOI:10.1016/j.chembiol.2003.09.002.

[98] L. M. Amzel, *Curr. Opin. Biotechnol.*, **1998**, DOI:10.1016/S0958-1669(98)80009-8.

[99] D. Ma, A. K. Bhattacharjee, R. K. Gupta, J. M. Karle, *Am. J. Trop. Med. Hyg.*,

1999, DOI:10.4269/ajtmh.1999.60.1.

[100] A. K. Bhattacharjee, R. K. Gupta, D. Ma, J. M. Karle, *J. Mol. Recognit.*, **2000**, DOI:10.1002/1099-1352(200007/08)13:4<213::AID-JMR500>3.0.CO;2-T.

[101] A. K. Bhattacharjee, W. Dheranetra, D. A. Nichols, R. K. Gupta, *QSAR Comb. Sci.*, **2005**, DOI:10.1002/qsar.200430914.

[102] J. B. Bhonsle, A. K. Bhattacharjee, R. K. Gupta, *J. Mol. Model.*, **2007**, DOI:10.1007/s00894-006-0132-0.

[103] R. García-Domenech, P. García-Mujica, Ú. Gil, J. M. Beltrán, J. Gálvez, *AfinidAd LXVii*, **2010**, *547*, 187–192.

[104] J. W. Pridgeon, J. J. Becnel, U. R. Bernier, G. G. Clark, K. J. Linthicum, *J. Med. Entomol.*, **2010**, DOI:10.1603/ME08265.

[105] R. Natarajan, S. C. Basak, A. T. Balaban, J. A. Klun, W. F. Schmidt, *Pest Manag. Sci.*, **2005**, DOI:10.1002/ps.1116.

[106] R. Natarajan, S. C. Basak, D. Mills, J. J. Kraker, D. M. Hawkins, *Croat. Chem. Acta*, **2008**, *81*, 333–340.

[107] R. García-Domenech, J. Aguilera, A. El Moncef, S. Pocovi, J. Gálvez, *Mol. Divers.*, **2010**, DOI:10.1007/s11030-009-9179-z.

[108] F. Tong, J. R. Coats, *Pest Manag. Sci.*, **2012**, DOI:10.1002/ps.3280.

[109] J. Song, Z. Wang, A. Findlater, Z. Han, Z. Jiang, J. Chen, W. Zheng, S. Hyde, *Bioorg. Med. Chem. Lett.*, **2013**, DOI:10.1016/j.bmcl.2013.01.015.

[110] E. Papa, F. Villa, P. Gramatica, *J. Chem. Inf. Model.*, **2005**, DOI:10.1021/ci050212l.

[111] L. Eriksson, J. Jaworska, A. P. Worth, M. T. D. Cronin, R. M. McDowell, P. Gramatica, *Environ. Health Perspect.*, **2003**, DOI:10.1289/ehp.5758.

[112] Y. Marrero-Ponce, M. Iyarreta-Veitía, A. Montero-Torres, C. Romero-Zaldivar,

C. A. Brandt, P. E. Ávila, K. Kirchgatter, Y. Machado, *J. Chem. Inf. Model.*, **2005**, DOI:10.1021/ci050085t.

[113] Y. Tang, W. Zhu, K. Chen, H. Jiang, *Drug Discov. Today Technol.*, **2006**, DOI:10.1016/j.ddtec.2006.09.004.

[114] M. Utsintong, T. T. Talley, P. W. Taylor, A. J. Olson, O. Vajragupta, *J. Biomol. Screen.*, **2009**, DOI:10.1177/1087057109344617.

[115] W. L. Jorgensen, *Science (80-.).*, **2004**, DOI:10.1126/science.1096361.

[116] P. D. Lyne, **2002**, *7*, 1047–1055.

[117] G. Schneider, H.-J. Böhm, *Drug Discov. Today*, **2002**, DOI:10.1016/S1359-6446(01)02091-8.

[118] Z. L. Liu, S. H. Ho, *J. Stored Prod. Res.*, **1999**, DOI:10.1016/S0022-474X(99)00015-6.

[119] B. Z. Sahaf, S. Moharramipour, M. H. Meshkatalsadat, *J. Asia. Pac. Entomol.*, **2008**, DOI:10.1016/j.aspen.2008.09.001.

[120] S.-W. Kim, J. Kang, I.-K. Park, *J. Asia. Pac. Entomol.*, **2013**, DOI:10.1016/j.aspen.2013.07.002.

[121] S. A. M. Abdelgaleil, M. I. E. Mohamed, M. E. I. Badawy, S. A. A. El-arami, *J. Chem. Ecol.*, **2009**, DOI:10.1007/s10886-009-9635-3.

[122] J. A. Pickett, M. A. Birkett, S. Y. Dewhirst, J. G. Logan, M. O. Omolo, B. Torto, J. Pelletier, Z. Syed, W. S. Leal, *J. Chem. Ecol.*, **2010**, DOI:10.1007/s10886-010-9739-9.

[123] Y. Huang, S. H. Ho, *J. Stored Prod. Res.*, **1998**, DOI:10.1016/S0022-474X(97)00038-6.

[124] Y. Huang, J. M. W. L. Tan, R. M. Kini, S. H. Ho, *J. Stored Prod. Res.*, **1997**, DOI:10.1016/S0022-474X(97)00009-X.

[125] D. P. Papachristos, D. C. Stamopoulos, *J. Stored Prod. Res.*, **2002**, DOI:10.1016/S0022-474X(01)00007-8.

[126] B. S. Tomova, J. S. Waterhouse, J. Doberski, *Entomol. Exp. Appl.*, **2005**, DOI:10.1111/j.1570-7458.2005.00291.x.

[127] S. Weaver, M. P. Gleeson, **2008**, DOI:10.1016/j.jmgm.2008.01.002.

[128] J. Jaworska, N. Nikolova-Jeliazkova, T. Aldenberg, *Altern. to Lab. Anim.*, **2005**, DOI:10.1177/026119290503300508.

[129] A. Tropsha, A. Golbraikh, *Curr. Pharm. Des.*, **2007**, DOI:10.2174/138161207782794257.

[130] E. Guenther, *Nature*, **1949**, DOI:10.1038/163663c0.

[131] M. Gütlein, A. Karwath, S. Kramer, *J Cheminform*, **2012**, *4*.

[132] Y. Kuwahara, W. S. Leal, Y. Nakano, Y. Kaneko, H. Nakao, T. Suzuki, *Appl. Entomol. Zool.*, **1989**, DOI:10.1303/aez.24.424.

[133] H. Yuan, R. He, B. Wan, Y. Wang, G. F. Pauli, S. G. Franzblau, A. P. Kozikowski, *Bioorg. Med. Chem. Lett.*, **2008**, DOI:10.1016/j.bmcl.2008.08.027.

[134] J. Yin, C. Wang, A. Mody, L. Bao, S.-H. Hung, S. A. Svoronos, Y. Tseng, *PLoS One*, **2013**, DOI:10.1371/journal.pone.0066598.

[135] M. A. Pathak, F. Daniels, T. B. Fitzpatrick, *J. Invest. Dermatol.*, **1962**, DOI:10.1038/jid.1962.106.

[136] Z. Korsak, J. Stetkiewicz, W. Majcherek, I. Stetkiewicz, J. Jajte, K. Rydzyński, *Int. J. Occup. Med. Environ. Health*, **2000**, *13*, 223–32.

[137] M. Humbert, S. Lavoine-hanneguelle, 2016.

[138] Y.-S. Hwang, K.-H. Wu, J. Kumamoto, H. Axelrod, M. S. Mulla, *J. Chem. Ecol.*, **1985**, DOI:10.1007/BF01024117.

[139] National Toxicology Program, *Natl. Toxicol. Program Tech. Rep. Ser.*, **1987**,

319, 1–198.

[140] G. Özek, Y. Suleimen, N. Tabanca, R. Doudkin, P. G. Gorovoy, F. Göger, D. E. Wedge, A. Ali, I. A. Khan, K. H. C. Başer, *Rec. Nat. Prod.*, **2014**, *8*, 242–261.

[141] K. Pålsson, T. G. T. Jaenson, P. Baeckström, A.-K. Borg-Karlson, *J. Med. Entomol.*, **2008**, DOI:10.1603/0022-2585(2008)45[88:TRSITE]2.0.CO;2.

[142] A. R. Bilia, F. Santomauro, C. Sacco, M. C. Bergonzi, R. Donato, *Evidence-Based Complement. Altern. Med.*, **2014**, DOI:10.1155/2014/159819.

[143] E. Innocent, C. C. Joseph, N. K. Gikonyo, M. H. H. Nkunya, A. Hassanali, *J. Insect Sci.*, **2010**, DOI:10.1673/031.010.5701.

[144] A. Khrimian, A. A. Cossé, D. J. Crook, *J. Nat. Prod.*, **2011**, DOI:10.1021/np200098z.

[145] E. Cavalieri, C. Bergamini, S. Mariotto, S. Leoni, L. Perbellini, E. Darra, H. Suzuki, R. Fato, G. Lenaz, *FEBS J.*, **2009**, DOI:10.1111/j.1742-4658.2009.07108.x.

[146] A. Khani, J. Asghari, *J. Insect Sci.*, **2012**, DOI:10.1673/031.012.7301.

[147] H. Chiong, T. L. Guggenheim, F. F. Khouri, M. L. Kuhlman, M. A. Navarro de Castro, R. R. Odle, B. A. Smith, 2013.

[148] D. Goel, R. Goel, V. Singh, M. Ali, G. R. Mallavarapu, S. Kumar, *J. Nat. Med.*, **2007**, DOI:10.1007/s11418-007-0175-2.

[149] W. Wanzala, A. Hassanali, W. R. Mukabana, W. Takken, *J. Parasitol. Res.*, **2014**, DOI:10.1155/2014/434506.

[150] M. Pal, A. Kumar, K. Tewari, *Facta Univ. - Ser. Physics, Chem. Technol.*, **2011**, DOI:10.2298/FUPCT1101057P.

[151] M. Makhaik, S. N. Naik, D. K. Tewary, *J. Sci. Ind. Res. (India).*, **2005**, *64*, 129–133.

[152] R. L. Rouseff, L. L. Stelinski, 2013.

[153] D. E. Wedge, J. A. Klun, N. Tabanca, B. Demirci, T. Ozek, K. H. C. Baser, Z. Liu, S. Zhang, C. L. Cantrell, J. Zhang, *J. Agric. Food Chem.*, **2009**, DOI:10.1021/jf802820d.

[154] C. Vongsombath, K. Pålsson, L. Björk, A.-K. Borg-Karlson, T. G. T. Jaenson, *J. Med. Entomol.*, **2012**, DOI:10.1603/ME12025.

[155] S. Cosimi, E. Rossi, P. L. Cioni, A. Canale, *J. Stored Prod. Res.*, **2009**, DOI:10.1016/j.jspr.2008.10.002.