# Analysis of the chemical space of antimalarial compounds by generative topographie mapping

Pavel Sidorov

# UNIVERSITÉ DE STRASBOURG

*ÉCOLE DOCTORALE DES SCIENCES CHIMIQUES*

**Chimie de la Matière Complexe – UMR 7140**

# THÈSE présentée par :

**Pavel SIDOROV**

soutenue le : **25 septembre 2017**

pour obtenir le grade de : **Docteur de l'Université de Strasbourg**

Discipline/ Spécialité : Chimie/Chémoinformatique

# Analyse de l'espace chimique des composés antipaludiques par la méthode GTM

**THÈSE dirigée par :**

| | |
|---|---|
| **Prof. VARNEK Alexandre** | Professeur, Université de Strasbourg |
| **Dr. HORVATH Dragos** | Directeur de recherche, Université de Strasbourg |

**RAPPORTEURS :**

| | |
|---|---|
| **Dr. TETKO Igor** | Docteur, Helmholtz Zentrum München |
| **Dr. MORELLI Xavier** | Directeur de recherche, Cancer Research Center of Marseille |

---

**AUTRES MEMBRES DU JURY :**

| | |
|---|---|
| **Prof. CAMPROUX Anne-Claude** | Professeur, Université Paris Diderot |
| **Prof. KELLENBERGER Esther** | Professeur, Université de Strasbourg |

**MEMBRES INVITES :**

| | |
|---|---|
| **Dr. DAVIOUD-CHARVET Elisabeth** | Directeur de recherche, Université de Strasbourg |

Pavel SIDOROV

# Analyse de l'espace chimique des composés antipaludiques par la méthode GTM


## Résumé

Cette thèse est consacrée à l'analyse de l'espace chimique des composés antipaludiques. L'analyse est faite à l'aide de la méthode des cartes topographiques génératrices (GTM). Un nouveau concept des cartes universelles est introduit et discuté en détail dans cette thèse : ce sont des cartes qui sont capables d'accommoder plusieurs jeux de données et les propriétés associées simultanément. Trois types des cartes sont construits et analysés : les cartes locales, globales et universelles. Elles sont toutes compétentes à la prédiction des composés actifs contre le parasite, ainsi qu'à l'analyse de l'espace chimique. Elles nous permettent d'étudier le recouvrement des données issues des sources différentes, de détecter des *terra incognita* de l'espace chimique, identifier des zones correspondantes aux différents mécanismes d'action, et révéler des incohérences d'annotations des données.


Mots-clés : espace chimique, visualisation, GTM, paludisme, cartes universelles.


## Résumé en anglais

This thesis is dedicated to the concept of the analysis of chemical space, and the application of that concept to antimalarial compounds. The analysis of the chemical space of antimalarial compounds here is done with the aid of the Generative Topographic Mapping (GTM) method. A concept of Universal GTM maps is developed and discussed in detail in this thesis: these are maps that are able to accommodate different datasets and associated properties. Three types of maps are built and analyzed: local, global, and universal. All these maps perform well in predicting compounds active against the parasite, as well as in the analysis of chemical space: they help us to study the overlap of data coming from different sources, detect *terra incognita* of the antimalarial space, delineate zones corresponding to various mechanisms of action, as well as highlight the inconsistencies in data annotations.


Key words : chemical space, visualization, GTM, malaria, universal maps

# UNIVERSITÉ DE STRASBOURG

*ÉCOLE DOCTORALE DES SCIENCES CHIMIQUES*

**Chimie de la Matière Complexe – UMR 7140**

# THÈSE présentée par :

**Pavel SIDOROV**

soutenue le : **25 septembre 2017**

pour obtenir le grade de : **Docteur de l'Université de Strasbourg**

Discipline/ Spécialité : Chimie/Chémoinformatique

# Analyse de l'espace chimique des composés antipaludiques par la méthode GTM

**THÈSE dirigée par :**

| | |
|---|---|
| **Prof. VARNEK Alexandre** | Professeur, Université de Strasbourg |
| **Dr. HORVATH Dragos** | Directeur de recherche, Université de Strasbourg |

**RAPPORTEURS :**

| | |
|---|---|
| **Dr. TETKO Igor** | Docteur, Helmholtz Zentrum München |
| **Dr. MORELLI Xavier** | Directeur de recherche, Cancer Research Center of Marseille |

---

**AUTRES MEMBRES DU JURY :**

| | |
|---|---|
| **Prof. CAMPROUX Anne-Claude** | Professeur, Université Paris Diderot |
| **Prof. KELLENBERGER Esther** | Professeur, Université de Strasbourg |

**MEMBRES INVITES :**

| | |
|---|---|
| **Dr. DAVIOUD-CHARVET Elisabeth** | Directeur de recherche, Université de Strasbourg |


# Analysis of the chemical space of antimalarial compounds

by

Pavel Sidorov
Laboratoire de Chémoinformatique
University of Strasbourg

A thesis submitted in September 2017
for the Degree of Docteur de lUniversite de Strasbourg

# *Abstract*

This thesis is dedicated to the concept of the analysis of chemical space, and the application of that concept to antimalarial compounds. The analysis of the chemical space of antimalarial compounds here is done with the aid of the Generative Topographic Mapping (GTM) method. A concept of Universal GTM maps is developed in discussed in detail in this thesis: these are maps that are able to accommodate different datasets and associated properties. Three types of maps are built and analyzed: local, global, and universal. All these maps preform well in predicting compounds active against the parasite, as well as in the analysis of chemical space: they help us to study the overlap of data coming from different sources, detect *terra incognita* of the antimalarial space, delineate zones corresponding to various mechanisms of action, as well as highlight the inconsistencies in data annotations.

# Declaration of Authorship

I, Pavel Sidorov, declare that this thesis titled, 'Analysis of the chemical space of anti-malarial compounds' and the work presented in it are my own. I confirm that:

- Information sources were acknowledged and fully referenced.
- I have acknowledged all main sources of help.

Pavel Sidorov

September 2017

# *Acknowledgements*

I thank my supervisors, Prof. Alexandre Varnek and Dr. Dragos Horvath, for having accepted me in the Laboratory of Chemoinformatics and having guided me during my PhD thesis. I am grateful as well to Dr. Gilles Marcou for his numerous and precious advises. I would like to express my gratitude to Dr. Héléna Gaspar, former PhD student in the laboratory, who has created the ISIDA/GTM software without which this work would be impossible.

I thank Timur Gimadiev and Dr. Kyrylo Klimenko for having provided the data that have been used in my work. I am very grateful to our colleagues: Dr. Birgit Viira from University of Tartu, Estonia, and Dr. Elisabeth Davioud-Charvet and Dr. Mourad Elhabiri from the Laboratory of Bioorganic and Medicinal Chemistry in Strasbourg, for their tremendous help with the data on antimalarial compounds.

I also thank everyone from the lab: Olga, thank you for your constant support; Grace, Fanny, Iuri, thanks for being such precious and generous friends; thank you also Fiorella, Olena, Julien, Arkadii, Marta, Shilva, my former intern Anne-Sophie, and there are surely others I forgot to mention.

And, of course, my family in Russia who did not see a lot of me these past 3 years, thank you for your support and faith in me.

# Contents

# Résumé en français

Le paludisme est une maladie infectieuse due à un parasite du genre *Plasmodium*, propagée par piqûre de moustique. Cette maladie grave est très répandue à l'Afrique sous-Sahari-enne, à l'Asie Sud-Est, et certaines parties de l'Amérique du Sud : elle cause 500 000 décès par an environ. Il existe des nombreux médicaments contre la maladie, dont l'efficacité est contrecarrée par l'émergence des souches résistantes du parasite. Par exemple, la chloroquine, le premier médicament synthétique produit dans les années 1930, a perdu rapidement son efficacité due à la résistance du parasite, et a été remplacé par d'autres composés, issus de différentes séries chimiques. Le traitement le plus utilisé à nos jours est la combinaison thérapeutique à base d'artémisinine (ACT). Néanmoins, la résistance à ces médicaments commence à se développer chez le parasite, surtout en Asie Sud-Est. Par conséquent, le besoin de nouveaux types de traitement visant d'autres voies métaboliques apparaît nécessaire.

Les études précédentes sur la conception des nouveaux médicaments antipaludiques ont exploré les relations structure-activité des petites séries de molécules limitées à une famille chimique. Seules les publications les plus récentes considèrent des jeux de données assez vastes, mais ces données requièrent des méthodes adaptées pour en extraire une information utile.

Il existe un nombre des médicaments antipaludiques en ce moment sur le marché. Une des familles les plus connues das composés antipaludiques est 4-aminoquinolines. Ces composés, dont le plus utilisé est la chloroquine, sont supposés d'agir sur le processus de la digestion de l'hémoglobine dans des érithrocytes. Artémisinine et ses dérivés sont aussi considérés d'intervenir dans la digestion, même si ce mécanisme précise est encore discuté. Un nombre des composés inhibitent le transport d'éléctrons dans le parasite: les exemples incluent les naphthoquinones (telles que l'atovaquone) qui inhibtient le cytochrome, ou les triazolopyrimidines, qui influencent la dihydroorotate déshydrogénase. Un autre mécanisme d'action intéressant est l'inhibition de la synthèse de l'acide folique. Cela se fait par un cycle compliqué des réactions, dont deux protéines

sont considérées comme des cibles potentielles: la dihydrofolate réductase et la dihydroptéroate synthase. Elles sont visées pas des différents médicaments, tels que cyclogunail, pyriméthamine, dapsone, sulfadoxine, etc. Des antiniotiques de large spectre (doxycycline, azithromycin) sont aussi bien utilisés dans des cas légers du paludisme, mais leur mode d'action n'est pas encore établi.

La complexité du paludisme est liée à ses nombreuses étapes impliquées dans le développement et la propagation de la maladie. Il existe un grand nombre des protéines qui sont considérées comme vitales pour le parasite, mais le mécanisme d'action et la cible précise de la majorité des médicaments antipaludiques n'est pas encore bien établi. Par conséquent, l'analyse des composés actifs contre le paludisme permettra, dans le cas idéal, de découvrir des nouveaux agents antipaludiques, mais aussi d'élucider leurs mécanismes d'action.

La thèse présente le développement d'une approche de l'analyse de l'espace chimique et de modélisation de plusieurs activités biologiques ou propriétés physico-chimiques par la méthode des Cartes Topographiques Génératrices (*Generative Topographic Mapping* en anglais, ou GTM), et son application à la modélisation de l'activité antipaludique. La thèse contient 7 chapitres. D'abord, nous introduisons la problématique du paludisme, des médicaments connus à nos jours et leurs mécanismes d'action, et la bibliographie sur la conception des composés actifs contre le paludisme *in silico*. Ensuite, la méthodologie générale de modélisation en chémoinformatique est décrite, ainsi que la méthode GTM utilisée dans ce travail. Le chapitre 4 couvre les données utilisées. Dans le cinqième chapître les développements méthodologiques sont expliqués. Ensuite, nous discutons leurs application à l'analyse de l'espace chimique des composés antipaludiques. Le chapître 7 contient une étude sur la modélisation des propriétés redox des naphthoquinones actives contre le parasite. Le dernier chapître présente un logiciel qui facilite l'analyse par GTM développé pendant la thèse.

## Données

Pour ce travail, nous avons utilisé deux sources principales de données. La majorité des données viennent de la base ChEMBL. La section ChEMBL-NTD (*Neglected Tropical Diseases*) contient plusieurs ensembles de données consacrés aux composés actifs contre variées maladies parasitaires, dont une grande partie est présentée par des composés antipaludiques. Deux grands jeux de données en ont été extraits : 1) TCAMS (Tres Cantos Antimalarial set), qui contient >13000 molécules avec l'activité antipaludique mesurée dans un seul essai ; 2) MalariaBox (Medicines for Malaria Venture PathogenBox), qui

contient 400 molécules dont l'activité est mesurée par plusieurs laboratoires suivant des méthodes variées.

Nous avons également utilisé les données obtenues par l'équipe de Dr. Davioud-Charvet à l'Université de Strasbourg. Il s'agit des activités antipaludiques mesurées dans plusieurs essais biologiques pour >200 molécules. Ces données ont été complétées par des composés de ChEMBL dont l'activité est établie selon des conditions expérimentales similaires. Dans les cas où les conditions expérimentales ont été différentes, nous avons utilisé une autre approche de fusion: trois médicaments connus on été sélectionnés pour jouer le rôle des composés de repère; si l'activité antipaludique pour ces trois était établie égale pour deux essais, il sont fusionnée. Ainsi, nous avons créé un jeu de donnée ActivityDB contenant 2093 molécules testées selon 17 protocoles expérimentaux différents. Pour chaque protocole, les molécules étudiées sont annotées actives ou inactives selon le seuil d'efficacité micromolaire.

La totalité des données issue de la fusion de TCAMS, MalariaBox et ActivityDB forme la base MalariaDB. Cette dernière contient 15462 molécules dont 1140 sont annotées par mécanisme d'action (expérimental ou hypothétique). Ces annotations viennent des publications originales (pour TCAMS et MalariaBox). Malheureusement, pas toutes les annotations sont robustes – on introduit le niveau de confiance: expérimental ou hypothétique. L'annotation est considérée hypothétique si elle n'a pas de confirmation expérimentale de la cible. Tel est le cas des données issues de TCAMS, où la cible biologique a été déterminée par similarité, par des études d'homologie, ou, au pires des cas, ce sont des cibles humaines.

Afin d'enrichir la collection, nous avons effectué une recherche supplémentaire dans la base CHEMBL. Cette recherche a été faite de manière automatique par des scripts écrit en Python, et a considéré un nombre de règles: la cible doit être annotée comme appartenant au *P. falciparum*, et le composé doit être actif contre elle. Cela nous a permis de trouver des confirmation pour une partie des données avec des annotations hypothétiques, ainsi que de trouver les annotations pour les composés non-annotées.

Un grand nombre des cibles ou des mécanismes d'action sont considérés, et souvent ils contiennent trop peu d'exemple pour être utile en modélisation. Nous avons alors catégorisé les cibles en groupes tenant compte de mode d'action ou de voie métabolique pour régrouper des cibles similaires. Au final, on a eu 8 catégories des mécanimes d'action: kinases, RCPGs, canaux ioniques, protéases, récepteurs nucléaires, cibles liées au transport d'électrons, digestion d'hémoglobine, ou glycolyse.

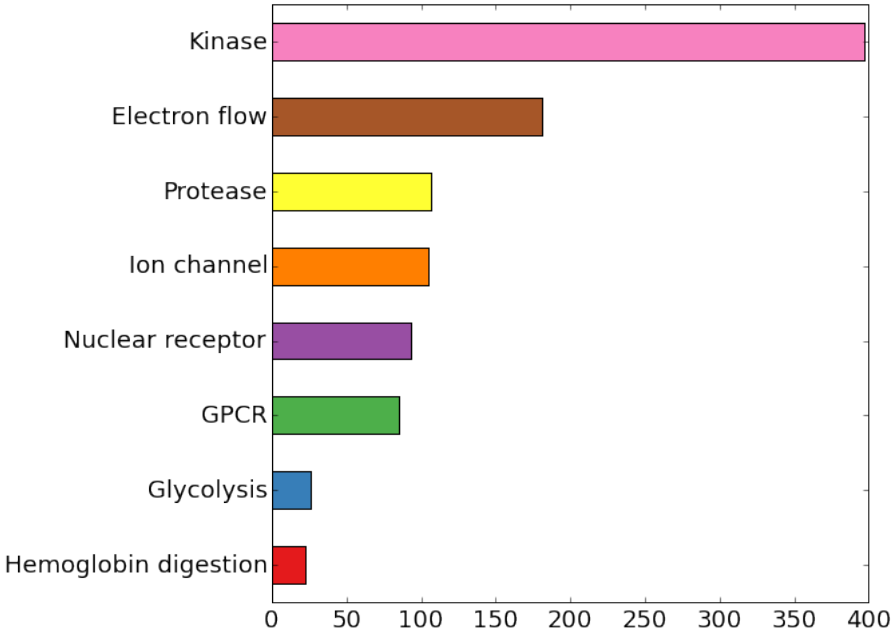



Figure 1. Nombre des molécules possédant le méchanisme d'action corrspondant à une catégorie des cibles considérées. Des cibles contenant très peu d'exemples ne sont pas indiquées.

## La méthode GTM

La méthode GTM (Cartes Topographiques Génératrices, *Generative Topographic Mapping*) est utilisée dans cette thèse pour la modélisation QSAR et l'analyse de l'espace chimique. La GTM prend le jeu de données encodé en $n$ descripteurs moléculaires (valeurs numériques décrivant la structure chimique) en entrée, $n$ est la dimension initiale. Une feuille flexible à deux dimensions (*la nappe*) s'adapte à ces données, et chaque molécule est projetée dessus. Quand la nappe est dépliée, on obtient une représentation de données en 2D dite une carte. Chaque composé a sa propre distribution de probabilité (la responsabilité) sur la carte.

Les vecteurs de responsabilité sont utilisé, en premier temps, pour visualiser la carte. Cela peut se faire de deux façons: 1) en calculant la moyenne pondérée de la resaponsabilité sur la carte, une molécule va être projetée dessus, donc sa location sur la carte sera définie; 2) en utilisant l'ensemble des molécules dont la propriété est connue, on pourra colorier la carte par la distribution de cette propriété.

Ensuite cette carte coloriée peut être utilisée pour prédire des propriétés des nouvelles molécules projetées dessus. La méthode GTM est capable de faire des modèles de régression et des modèles de classification. Dans le cas des modèles de classification, deux options sont présentes: soit colorier les noeuds de la carte par la classe dominantes, c'est-à-dire, celle-là qui possède les plus exeples qui contribue à ce noeud; soit la coloration

se fait par la propriété moyenne de toutes les molécules qui sontribuent dans chaque noeud. Cette dernière façon est aussi applicable dans le cadre des modèles de régression. La prédiction se fera soit en utilisant toue la carte, soit seulement les noeuds les plus proches.

Une autre façon d'utiliser des vecteurs de responsabilité est pour l'analyse de l'espace chimique grâce aux motifs de responsabilité. Les motifs de responsabilité est une méthodologie de régroupement des composés similaires sur une carte, ainsi utilisée pour trouver des zones peuplées par des molécules possédant le même châssis moléculaires.

L'évaluation des modèles construits à l'aide de la méthode GTM est basée sur les paramètres statistiques classiques utilisés en chémoinformatique. Pour les modèles de régression, on utilise les paramètres $RMSE$ (*Root Mean Square Error*, la racine de l'erreur carrée moyenne de la prédiction) et $R^2$ (coefficient de détermination). Ils sont calculés selon les formules suivantes:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N} (y_{exp,i} - y_{pred,i})^2}{N}}. \tag{1}$$

Ici, $N$ est la taille du jeu de données considéré, $y_{exp,i}$ et $y_{pred,i}$ les valeurs expérimentales et prédites par le modèles de la propriété de la molécule numéro $i$.

$$R^2 = 1 - \frac{\sum_{i=1}^{N} (y_{exp,i} - y_{pred,i})^2}{\sum_{i=1}^{N} (y_{exp,i} - < y_{exp} >)^2}, \tag{2}$$

$< y_{exp} >$ signifie la valeur moyenne de la propriété pour toutes les molécules du jeu de données.

L'évaluation des modèles de classification se fait différemment. La plupart des tâches de classification sont binaire, donc deux classes d'entités sont présentes. Dans ce cas-là, on construit une matrice de confusion contenant la quantité des entités de chaque classe predit correctement et incorrectement:

| | | Résultat de prédiction | |
|---|---|---|---|
| | | Classe 1 | Classe 2 |
| Valeur réelle | Classe 1 | Vrai positif ($TP$) | Faux négatif ($FN$) |
| | Classe 2 | Faux positif ($FP$) | Vrai négatif ($TN$) |

A partir de ces nombres, on peut calculer les charactéristiques numériques de la qualité du modèle. L'une que l'on utilise dans ce travail est *Balanced accuracy (BA)* qui se calcule par la formule suivante:

$$BA = \frac{1}{2}\left(\frac{TP}{TP+FN} + \frac{TN}{FP+TN}\right). \quad (3)$$

Ce paramètre est surtout utilisé pout les modèles construits sur les jeux de données mal équilibrés.

Pour estimer ces valeurs, on utilse aussi la technique de la validation croisée (*cross-validation*). Elle permet d'évaluer le modèle sans utiliser des données externes. Pour ce faire, on découpe le jeu de données initial en plusieur paquets. On utilisera donc un paquet pour la validation et le reste – pour l'entraînement du modèle. On répétera la procédure en changeant le paquet pour que à la fin chaque molécule du jeu de données initial soit utilisée pour la validation une fois. L'évaluation se fera avec des valeurs prédites dans la validation.

## Développements métodologiques

La méthode GTM a été précédemment utilisée avec succès pour la modélisation et la prédiction des propriétés individuelles (numériques ou catégoriques) des molécules. Dans ces études, elle a été limitée, comme la majorité des méthodes d'apprentissage automatique, à une propriété à la fois. Par contre, les particularités de la méthode ouvrent la possibilité de l'utiliser afin de modéliser plusieurs activités simultanément. Le point le plus important est que la construction d'un modèle GTM se fait en deux étapes : la première est l'ajustement de la nappe aux données de façon non-supervisée (ignorant les propriétés des molécules) ; la deuxième est la "coloration" de la carte par la propriété du jeu d'entraînement et la prédiction d'un jeu externe. Grâce à la séparation de ces étapes, nous pouvons colorer une même carte en utilisant plusieurs jeux de données avec des propriétés différentes.

Le modèle GTM dépend, évidemment, des paramètres de la méthode, de l'ensemble de molécules sur lequel il est construit et des descripteurs moléculaires choisi. Il est donc nécessaire de développer une stratégie de sélection de tous ces paramètres afin de créer une carte capable de prédire plusieurs propriétés simultanément.

La stratégie proposée dans ce projet est la suivante (Figure 2). En entrée, le jeu de donnée initial (dit *frame set*), les descripteurs moléculaires et les paramètres de la méthode sont donnés. La nappe est ajustée aux données et définit l'espace chimique du modèle. Ensuite, les jeux externes (*selection sets*) sont projetés un par un sur cette carte. Chaque jeu est évalué dans un modèle de régression ou de classification en validation croisée, et le score moyen des performances des modèles obtenus détermine la qualité de la carte.

Un algorithme génétique est utilisé pour échantillonner efficacement toutes les combinaisons possibles de ces paramètres d'entrée. Les meilleures cartes sont celles qui sont les plus performantes pour la modélisation des propriétés des jeux externes. Elles ont, donc, les meilleurs scores, et peuvent être exposées à une validation supplémentaire.

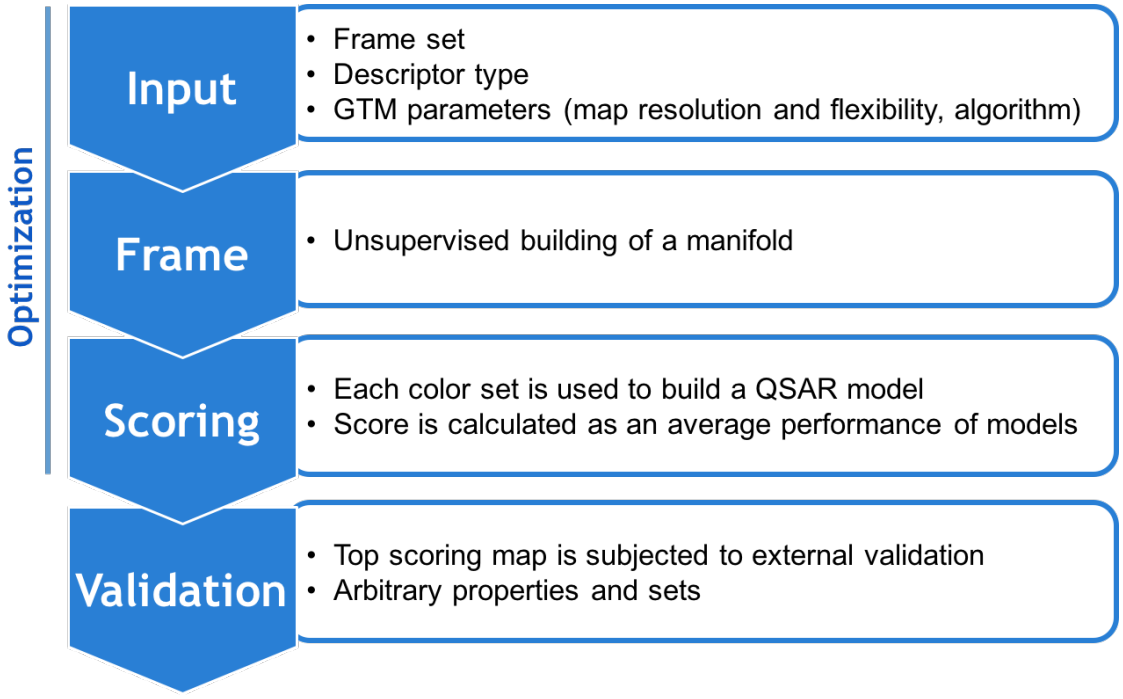



Figure 2. Représentation schématique de la stratégie d'optimisation des paramètres de la carte GTM à la base des algorithmes génétiques.

Cette stratégie a été testée sur un jeu de données extrait de ChEMBL (fourni par Prof. J. Bajorath, Université de Bonn, Allemagne). Il s'agit des inhibiteurs des 144 protéines humaines (l'activité expérimentale est mesuré *in vitro*), soit plus de 30 000 composés uniques au total, qui ont joué le rôle des jeux de sélection.

Les frame sets incluent donc les échanitllon du jeu de données décrit au-dessus, ainsi que des médicaments connus, des molécules des bases PubChem et ZINC, en différentes proportions. Au total, 5 frame sets différents on été considérés. Quant aux descripteurs moléculaires, 39 types de fragments ISIDA ont été utilisés.

Cinq meilleures cartes on été sélectionnées pour l'analyse et la validation externe. 450 jeux des données supplémentaires concernant les inhibiteurs des autres cibles biologiques humaines, les protéines transporteurs, ainsi que des composés antipaludiques et antiviraux, ont été issues de ChEMBL pour une validation externe. Les cartes performent bien en prédiction des molécules actives contres les protéines humaines et des composés antiviraux. Elles séparent bien les molécules actives et inactives contre une cible individuelle, ainsi que les familles des cibles (kinases, protéases, etc.) entre elles. En plus,

les cartes ont des bonnes performances sur des propriétés plus compliquées – activités antivirales. Un exemple est présenté à la figure ci-dessous.

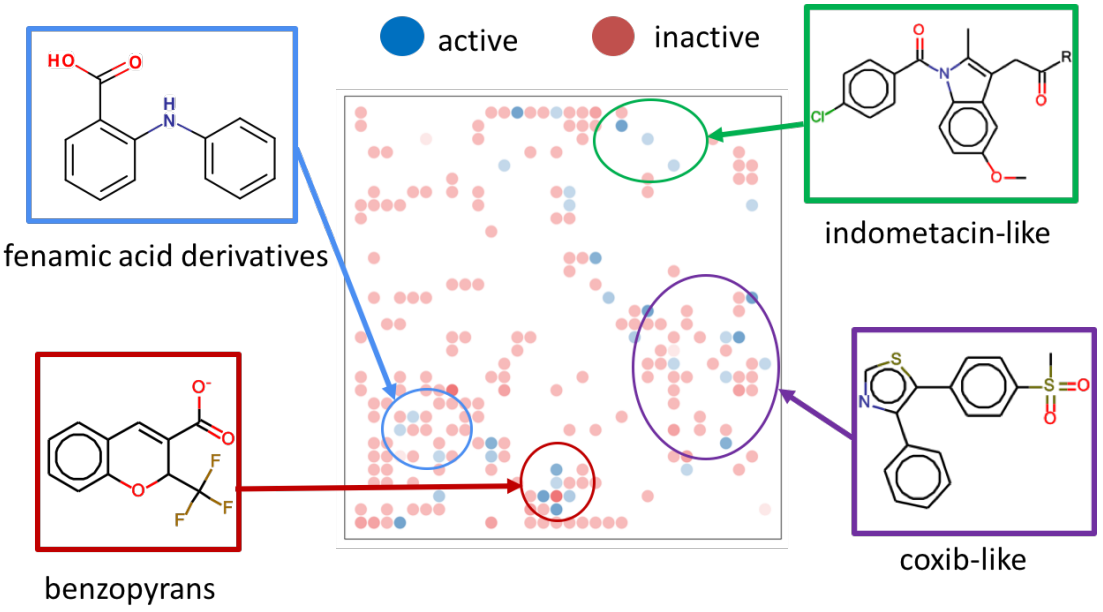



Figure 3. L'exemple d'un jeu de données des ligands de cyclooxygénase 2 (Cox-2) sur la carte universelle. Le BA en validation croisée est égal à 0.7. La coloration est faite par la classe dominante (actif/inactif). Dans les zones indiquées - les châssis communs des molécules. Comme on peut voir, des différentes familles des inhibiteurs de Cox-2 sont bien séparées sur la carte.

Il a été aussi intéressant pour nous de voir le recouvrement des différents jeu de données sur les cartes obtenues. Nous avons étudié la récouvrement des sous-jeux concernant différentes familles des protéines humaines. Nous avons trouvé, que des jeux de données des RPCGs sont en général specifiques à la protéine, donc le recouvrement des jeux dédiés aux différentes protéines est minimal. D'autres côté, les inhibiteurs des kinases se trouvent souvent bien dispersé sur la carte et indépendant du type de la kinase. Cela veut dire que es inhibiteurs des kinases sont en geénéral pas spécifiques, ce qui est confirmé par expérience.

En conclusion, on peut dire que l'approche de la carte universelle performe bine pour la modélisation multi-tâche. Les cartes obtenues performent bien en régression et classification d'une variété des activité biologiques, et nous aident à analyser l'espace chimique des protéines et des familles des protéines. Cela nous permet de les considérer comme des cartes universelles de l'espace chimique des médicaments.

## Cartes de l'espace des composés antipaludiques

Les 5 meilleures cartes "universelles" choisies selon leurs bonne performance sur des jeux de sélection et jeux de validation externes n'ont pas eu beaucoup de succès à la prédiction des composés antipaludiques, donc il a été décidé de créer des cartes déidées à l'espace chimique des composés antipaludiques. Pour ce faire, deux stratègies ont été suivi.

Afin de pouvoir construire un modèle général pour l'espace chimique des composés antipaludiques qui prendrait en compte l'activité et le mécanisme, le jeu de données MalariaDB a été utilisé. La stratégie d'optimisation d'une carte par l'algorithme génétique est appliqué à ce jeu. La fonction optimisée est la performance moyenne sur classification des mécanismes d'action. Quatre meilleures cartes ("globales") ont été sélectionnées pour l'analyse.

D'autre côté, une carte "locale" a été construite spécifiquement sur les données ActivityDB pour maximiser la performance de prédiction de la classe d'activité antipaludique des molécules.

Afin d'évaluer la qualité des cartes obtenues au niveau de leur pouvoir prédictif, les molécules de l'ActivityDB ont été projetées sur les cartes "universelles" (construites sur les données ChEMBL) d'une part, et sur une carte "locale" et les cartes "globales" d'autre part. Tous les types de cartes séparent bien les actifs et les inactifs dans les 17 sous-ensembles de données correspondant aux différents protocoles : au moyen, Balanced Accuracy BA est égale à 0.8 pour la carte locale, 0.74 pour la carte universelle, et autour 0.74 pour les cartes globales. Cette performance est comparée à la performance des modèles SVM construits précédemment sur les mêmes données: le BA moyen des modèles SVM est égal à 0.81. Il faut se dire en même temps que la façon de la construction des modèles est différente pour ces deux méthodes. Les modéles SVM sont faits individuellement pour chaque jeux de donnés, et l'optimisation des paramètres et des descripteurs mène à un ensemble des paramètres spécifique pour chaque modéle. La GTM, au contraire, crée un seul modèle – une carte – qui peut être coloriée par des propriétés pour les prédire pour des nouvelles molécules. Le fait que la performance du modèle GTM est en même niveau que la performance des modèles classiques SVM confirme bien le pouvoir de la méthode GTM et son applicabilité dans le cadre des modèles multi-tâches.

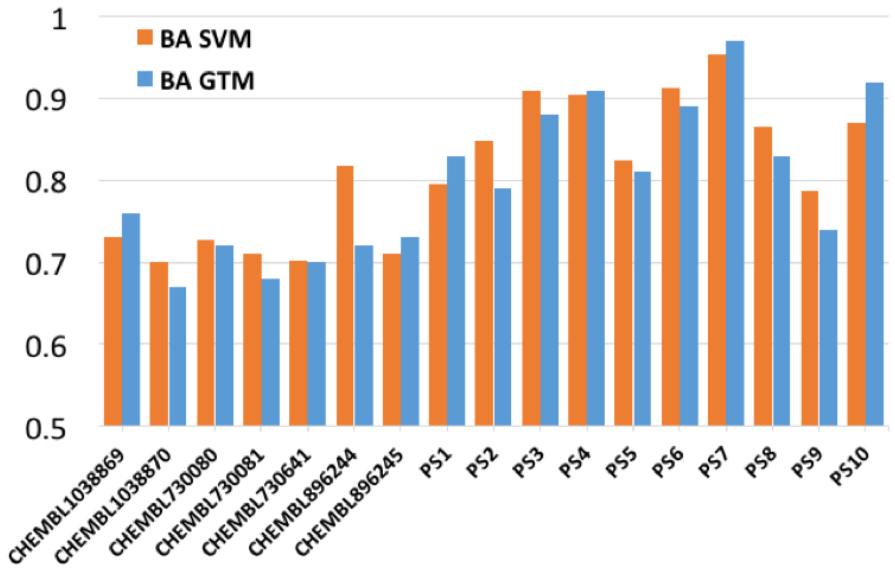



Figure 4. Comparaison de la performance des modéles GTM (carte locale) et des modéles SVM individuels construits sur les jeux de données de l'ActivityDB. Les modéles GTM sont en général aussi performants que les modèles SVM malgré la différence des façons de construction des modéles.

## Analyse de l'espace chimique

Le point le plus intéressant de ce travail est néanmoins l'application de la méthode GTM pour l'analyse de l'espace chimique. L'analyse de l'espace chimique se base sur les cartes globales et locale, car elles sont les plus pertinentes pour l'activité antipaludique.

Prenons d'abord l'exemple des cartes globales. La détection des motifs de responsabilité nous a permis de trouver des châssis communs pour des mécanismes d'actions distincts (tel que les inhibiteurs de kinases ou du cytochrome), et encore, distinguer des zones correspondant aux différentes cibles biologiques. Par exemple, dans le cas des inhibiteurs du transport d'électrons, on peut voir des zones séparées pour les inhibiteurs du cytocrome et les inhibiteurs de la dihydroorotate déshydrogénase. Ce qui est intéressant c'est que ces motifs sont en général transferts entre des cartes, donc on peut trouver des mêmes motifs sur chaque carte. Bien sûr, certaines cartes, par exemple la carte 2 qui est basée sur des descripteurs du type pharmacophorique, nous permettent des trouver davantage l'information spécifique aux structures qui diifèrent néanmoins de point de vue chimique, mais pas biologique , ce qui peut être intéressant pour la découverte des nouveaux chémotypes.

Nous avons pu aussi découvrir certaines incohérences dans les annotations. Dans certaines zones compactes, nous avons détecté les molécules voisines annotées différemment. Par exemple, nous avons détecté des composés de la famille des 4-aminoquinolines, qui sont en général considérés d'interrompre la digestion de l'hémoglobine, dans une

zone peuplée par des inhibiteurs de kinases. Cela peut nous permet de suggérer un mécanisme d'action supplémentaire pour ces molécules.

Un autre example est une zone d'une population mixte: des molécules avec des annotations hypothétiques et expérimentales résident dans le même endroit. Dans ce cas, si l'une de ces molécules a l'annotation expérimentale et d'autres ont les annotations hypothétiques, ces dernières doivent être corrigées. Cette stratégie a été utilisée corriger l'annotation erronée de certaines molécu-les (voir des exemples sur la Figure 5).

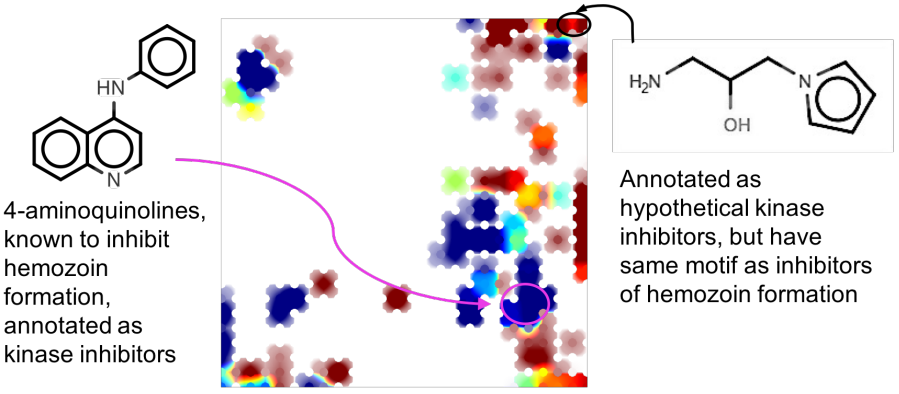



Figure 5. La carte globale (Map2) construite sur MalariaDB et coloriée par le mécanisme d'action (bleu pour inhibiteurs de kinases, rouge pour autres mécanismes, les couleurs intermédiaires signalent la population mixte dans les nœuds). Deux exemples de motifs structuraux correspondent à des incohérences d'annotation par cible, comme indiqué.

L'application très importante de cette approche est la possibilité de suggérer des modes d'action pour des nouvelles molécules. Vu que la base MalariaDB contient une grande quantité des composés non-annotés, il serait d'un grand intérêt de faire un criblage virtuel de cette base pour prédire les cibles, voire corriger ou proposer des nouveaux mécanismes d'action pour les molécules annotées.

La carte locale, d'autre côté, nous permet d'étudier le recouvrement des jeux de données impliqués dans ActivityDB, ainsi que de détecter les motifs structuraux les plus et le moins investigués. Figure 6 montre la carte locale avec des zones qui correspondent à certains motifs stucturaux. Par exemple, le centre de la carte (zone 1) est peuplé par des composés de la série de 4-aminoquinolines, la famille des antipaludiques la plus étudiée. On peut aussi voir des zones des artémisinins, quinones, etc. Cela est important pour prouver que la carterespecte le principe des voisins – les molécules similaires sont localisées proche l'une de l'autre.

Au contraire, des zones isolées peuplées par des molécules actives possédantes un motif structural rare, qui constituent donc *terra incognita* de l'espace chimique antipaludique, sont très intéressants et nécessitent d'être étudiées. Par exemplé, plusiers "îles" isolées sont trouvées sur la carte. Ces îles contiennent un nombre des composés actifs, et donc

il est important de décrire l'espace chimique autout d'elles. De ce fait, les cartes peuvent guider des chimistes dans leur recherche des nouveaux châssis prometteurs pour l'activité antipaludique.

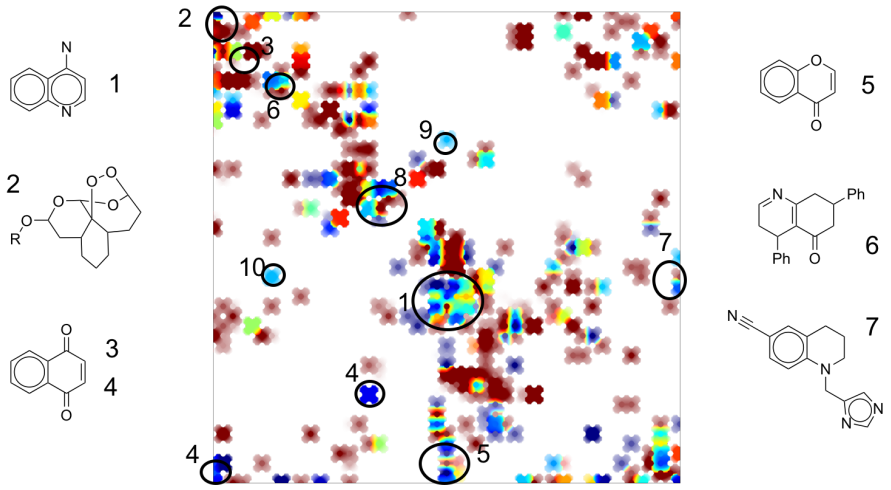



Figure 6. La carte locale des molécules antipaludiques. Les nœuds de la carte sont coloriés en bleu s'ils sont peuplés majoritairement par des composés actifs, sinon en rouge. Des zones associées à certains motifs structuraux privilégiés sont indiquées sur la carte.

## Modélisation des propriétés redox

L'équilibre oxydo-réductif est important pour la vie du parasite, surtout à l'étape hématique. Nos collaborateurs du Laboratoire de chimie bioorganique et médicinale nous a fourni l'information sur la série des 1,4-naphthoquinones substituées. Nous l'avons ensuite complétée par des données de la littérature (en prenant des quinones et indolone-N-oxides dont l'activité antipaludique a été aussi étudiée) afin de former un jeu de données de 95 molécules qui a été soumis à la modélisation de la propriété désirée – le potentiel redox mesuré en voltamètrie cyclique.

Deux types de descripteurs moléculaires ont été utilisés – les descripteurs fragmentaux ISIDA du type séquence d'atomes et liaisons de longueur 2 à 10 atomes et les descripteurs d'effet électronique (EED).

Les descripteurs EED ont été développés au cours de ce travail. Vu que la propriété qu'on cherche à modéliser ici – le potentiel redox – est très dépendant de la structure éléctronique de la molécule, il semble raisonable de baser les descripteurs dessus. Les descripteurs EED décrivent l'influence de l'environnement d'un atome marqué du point de vue de sa structure électronique. Les propriété de la structure éléctronique sont calculées avec le paquet des logiciels ChemAxon. Ce sont, par exemple, la charge partielle

sur les atomes, conjugaison, état d'hybridisation, polarisabilité, densité électronique, l'énergie nucléophilique ou éléctrophilique, etc.). Ces propriété sont pondérées à la base de distance topologique entre le groupe marqué – dans ce cas-là, le groupe carbonyl dans l'anneau quinone – et les atomes voisins. La distance maximale de 4 liaisons est imposée. Au total, 704 descripteurs sont calculés pour chaque molécules.

Le jeu de données a été divisé en deux parties – le jeu d'entraînement (81 composés) et le jeu de test (14), afin de pouvoir valider le modèle en validation croisée et par le test externe. Les méthodes de la régression multilinéaire (MLR, le logiciel ISIDA/QSPR) et des vecteurs supports (SVM, la librairie libsvm) ont été appliquées. Dans le cas des modèles MLR, un modèle consensus a été construit sur tous les types des descripteurs considérés, donc la prédiction des nouvelles molécules se base sur la moyenne des prédiction de chaque modèle. Dans le cas de SVM, de l'autre côté, le modèle a été optimisé pour la prédiction le plus précise.

Les modèles obtenus ont une bonne performance en validation croisée et en validation externe. Les valeurs du potentiel redox du test (expérimentales contre prédites) sont présentées sur la figure 3. Un point aberrant du modèle EED est dû aux limitations des descripteurs. Le modèle consensus MLR sur les descripteurs ISIDA marche généralement mieux (l'erreur moyenne de prédiction sur le jeu externe est 0.032 contre 0.054 en EED) et est plus interprétable de point de vue chimique. La moins bonne précision du modèle sur les descripteurs EED s'explique néanmoins par la présence d'un point aberrant: un composé avait un groupe hydroxyle éloigné du centre marqué, donc cela n'a pas pu être pris en compte par le modèle. Ces modèles sont accessibles en ligne dans l'outil du prédiction du laboratoire (`infochim.u-strasbg.fr/webserv/VSEngine.html`).

## Developpement des logiciels

Les outils de la visualisation des cartes GTM dévéloppés précédemment produisait seulement des images statiques. Afin d'aider l'utilisateur à analyser l'espace chimique par une carte GTM, un outil interactif de visualisation a été développé en Javascript à l'aide de la librairie D3. Cet outil permet d'entrer l'information sur la carte obtenue (la nappe) et les molécules (identificateur ChEMBL, structure en format SMILES, label d'intérêt, etc). Il visualise la carte, colorie les noeuds par des classes indiquées et projette des molécules dessus. Les éléments interactifs permettent de sélectionner et visualiser des molécules sur la carte. Quand l'utilisateur sélectionne une molécule, il voit la structure et toutes les informations que le créateur de la carte avait fournies sur l'entrée au tableau à droite.

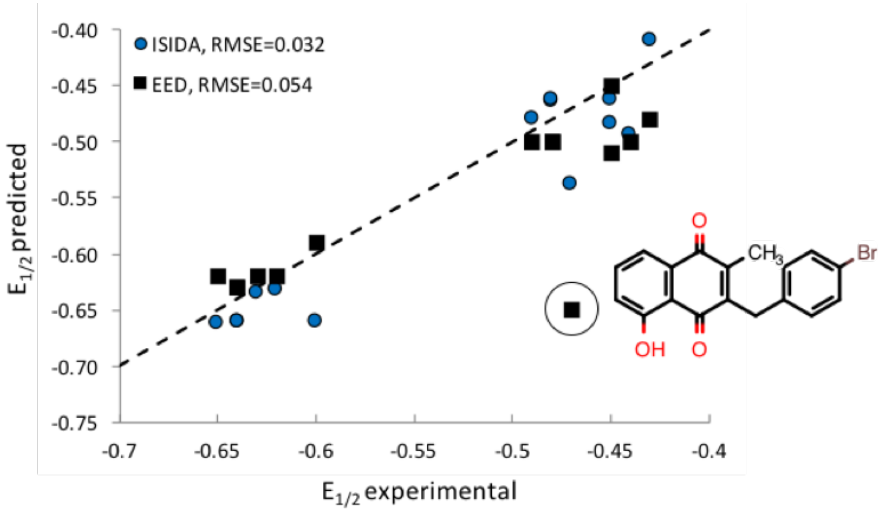



Figure 7. Potentiel redox expérimental (axe X) et prédit (axe Y) par les modèles utilisant des descripteurs ISIDA (•) et EED (■) du jeu de test de naphthoquinones substitués. Le point aberrant de prédiction est entouré. La ligne pointillée représente la prédiction idéale.

Afin de simplifier l'analyse des composés voisins une option lasso a été aussi introduite dans l'outil.

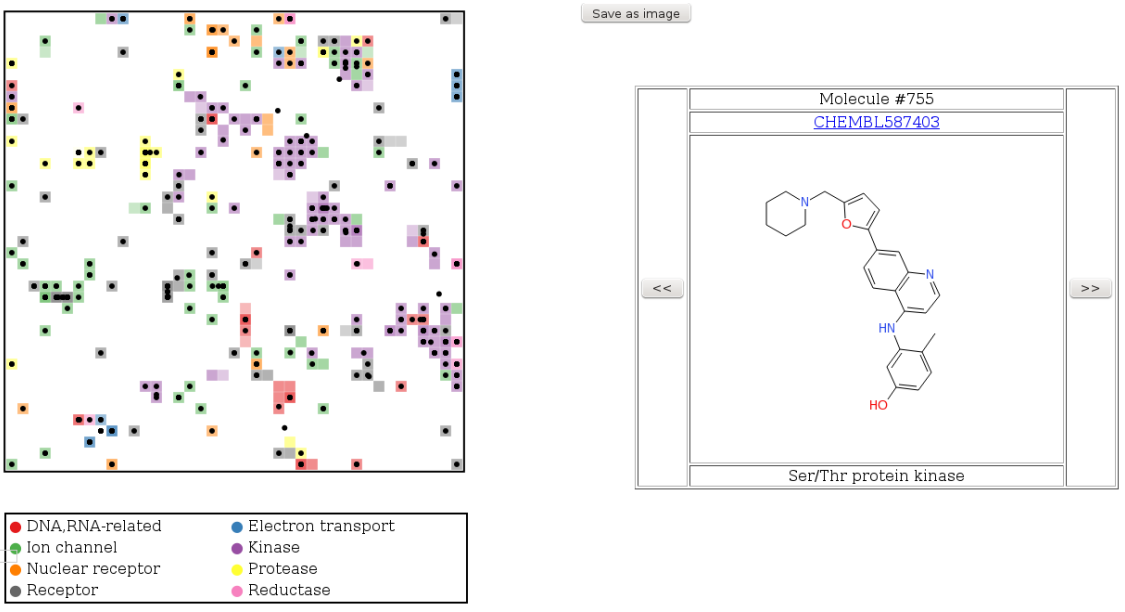



Figure 8. L'interface graphique de l'outil de visualisation des cartes GTM. La carte est à gauche, l'information sur des molécules sélectionnées est à droite.

## Conclusion générale

Cette thèse est consacrée à l'analyse de l'espace chimique des composés antipaludique. Le problème de la conception des nouveaux médicaments antipaludiques est très important en vue d'émergence des souches résistantes du parasite. Afin de développer des modèles prédictifs par des méthodes *in silico*, un grand jeu de données des molécules antipaludiques avec leurs activités mesurées et, si possible, le mécanisme d'action a été collecté et nettoyé soigneusement. Il contient 15462 molécules, dont 1140 sont annotées par leur cible biologique. 2093 entre elles sont annotées par leur activité antipaludique, mésurée dans un de 17 essai biologiques.

La complexité des données sur les molécules antipaludiques est liée à leur hétérogénéité. Il est difficile des fusionner des données issués des différentes sources, donc des méthodes classiques de l'apprentissage automatique seront défavorisé dans ce cas. Pour cela, la méthode des cartes topographiques génératrices (GTM) a été appliquée. Une stratégie de paramétrisation de la méthode GTM a été proposée. Cette stratégie permet de sélectionner les meilleurs descripteurs et paramètres pour construire une carte adaptée à la prédiction de plusieurs propriétés simultanément. Elle a été testée sur une « carte universelle » : une carte capable de prédire l'activité des molécules contre 144 protéines humaines. Elle se généralise sur les jeux de données externes : activité contre plus de 400 protéines (qui ne participaient pas dans la paramétrisation) et les activités des composés antipaludiques et antiviraux *in vivo*. plus, les cartes peuvent être utilisées pour investiguer l'espace chimique et la distribution des chémotypes sur la carte.

Cette stratégie a été appliquée sur le jeu ActivityDB contenant les molécules dont l'activité antipaludique est mesurée (2093 molécules). La meilleure carte locale fonctionne très bien dans la classification des composés actifs et inactifs, les résultats sont comparables aux modèles construits sur le même jeu de données par la méthode SVM. L'avantage de la carte est la visualisation : une représentation 2D permet de distinguer des zones de molécules actives contre une cible biologique.

La carte globale de l'espace chimique des composés antipaludiques a été construite sur le jeu de données entier (15462 molécules). Elle sépare bien les molécules agissant selon des mécanismes différents et discrimine les composés actifs et inactifs avec une bonne performance. En plus, cette carte nous a permis de détecter et de corriger certaines incohérences d'annotations des mécanismes d'action et suggérer des nouveaux mécanismes d'action.

Le mécanisme d'action des antialudiques par les cycles redox dans le parasite a été étudié en détail. Pour construire des modéles prédictifs, un nouveau type de descripteurs moléculaires locaux moléculaires – les descripteurs d'effet électronique (EED) – a été

proposé. Ces descripteurs décrivent l'influence de l'environnement d'un atome marqué du point de vue de sa structure électronique (charge partielle sur les atomes, conjugaison, etc.). Un modèle QSAR de prédiction des propriétés redox des naphthoquinones a été construit en utilisant les descripteurs ISIDA et EED. Les modèles développés sont accessibles pour les utilisateurs en-ligne.

Enfin, pour faciliter l'analyse des cartes GTM, un outil interactif a été développé. Cet outil permet de visualiser la carte, ainsi que parcourir les structures des molécules en temps réel.

## Posters

- **27 juin – 1 juillet 2016**

  Strasbourg Summer School in Chemoinformatics 2016 (Strasbourg, France), ***Computational investigation of phenotypic anti-malarial compounds***, P. Sidorov, B. Viira, D. Horvath, G. Marcou, U. Maran, E. Davioud-Charvet, A. Varnek.

- **27 juin – 1 juillet 2016**

  Strasbourg Summer School in Chemoinformatics 2016 (Strasbourg, France), ***Visualization and analysis of large dataset of chemical reactions using GTM***, P. Sidorov, A. Lin, T. Madzhidov, A. Varnek.

- **27 juin – 1 juillet 2016**

  Strasbourg Summer School in Chemoinformatics 2016 (Strasbourg, France), ***Chemical reactions visualization: do outliers make any sense?*** , I. Casciuc, P. Sidorov, A. Varnek.

- **4-8 septembre 2016**

  21st EuroQSAR Symposium (Vérone, Italie), ***Visualization and analysis of the chemical space of antimalarial compounds***, P. Sidorov, B. Viira, D. Horvath, G. Marcou, E. Davioud-Charvet, A. Varnek.

- **26-30 septembre 2016**

  XX Mendeleev congress on general and applied chemistry (Ekaterinburg, Russie), ***Chemical space of hydrogenation reactions: visualization and analysis***, A. Lin, P. Sidorov, T. Madzhidov, A. Varnek.

## Communications orales

- **6-9 juillet 2015**

  Second Kazan Summer School on Chemoinformatics (Kazan, Russie), ***Mappability of Drug-like Space: towards a polypharmacologically competent map of drug-relevant compounds***, P. Sidorov, H. Gaspar, G. Marcou, D. Horvath, A. Varnek.

- **8-9 octobre 2015**

  7èmes journées de la SFCi (Nice, France), ***Mappability of drug-like space***, P. Sidorov, H. Gaspar, G. Marcou, D. Horvath, A. Varnek.

- **14 janvier 2016**

  Journée scientifique de l'UMR 7140 (Strasbourg, France), ***Towards a universal map of drug-like chemical space***, P. Sidorov, D. Horvath, A. Varnek.

## Publications

- Elhabiri M, Sidorov P, Cesar-Rodo E, Marcou G, Lanfranchi D A, Davioud-Charvet E, Horvath D, Varnek A. **Electrochemical Properties of Substituted 2-Methyl-1,4-Naphthoquinones: Redox Behavior Predictions**. *Chemistry – A European Journal*, 21 (8), pp. 3415-3424, 2015.

- Sidorov P, Gaspar H, Marcou G, Varnek A, Horvath D. **Mappability of drug-like space : towards a polypharmacologically competent map of drug-relevant compounds**. *Journal of Computer-Aided Molecular Design*, 29 (12), pp. 1087-1108, 2015.

- Sidorov P, Desta I, Chessé M, Horvath D, Marcou G, Varnek A, Davioud-Charvet E, Elhabiri M. **Redox Polypharmacology as an Emerging Strategy to Combat Malarial Parasites**. *ChemMedChem*, 11 (12), pp. 1339-1351, 2016.

- Gaspar H, Sidorov P, Horvath D, Baskin I, Marcou G, Varnek A. **Generative Topographic Mapping approach to chemical space analysis**, *Frontiers in Molecular Design and Chemical Information Science-Herman Skolnik Award Symposium 2015: Jürgen Bajorath*, pp. 211-241, 2016.

- Sidorov P, Viira B, Davioud-Charvet E, Maran U, Marcou G, Horvath D, Vanek A. **QSAR modeling and chemical space analysis of antimalarial compounds**, *Journal of Computer-Aided Molecular Design*, 31 (5), pp. 441-451, 2017.

# Chapter 1

# Introduction

Malaria (the term originates from Medieval Italian: *mala aria* – "bad air") remains one of the most severe infectious diseases caused by parasitic protozoans (a group of single-celled microorganisms) belonging to the *Plasmodium* type. According to the WHO, 214 million clinical cases of malaria have been reported worldwide in 2015. While a 48% decline in mortality since 2000 has been observed, malaria is still the fourth highest cause of death.

Chapter 2 gives an overview on the state-of-the-art of the antimalarial drug design. While there is a plethora of drugs that have been developed to treat malaria since the beginning of the 20th century, for most of them, the precise target is still not known. The emergence of drug resistance in the parasite causes the loss of efficacy even for the combination therapies. Another problem of many drugs against malaria is their cost, which limits the access to them in the regions that are most in need. Therefore, the search for new efficient antimalarial chemical entities continues.

*In silico* methods help to accelerate and facilitate the discovery and design of novel active molecules. Quantitative Structure-Activity Relationship (QSAR) modeling is one of the major computational tools employed in drug design and discovery. Chapter 3 describes the general practices of QSAR modeling, and discusses in more detail the method that has been used during the thesis – Generative Topographic Mapping (GTM). GTM combines the predictive power of other machine learning methods with the visualization aspect that is a practical way to convey useful and interpretable information for a specific purpose. It a very useful tool for the analysis of chemical space. In this thesis, the chemical space is described as a $D$-dimensional descriptor space, where molecules are characterized by $D$ chemical descriptors. Such analysis helps in identifying dense or sparse zones of chemical space which can be used for comparing datasets, highlighting missing or dominating chemotypes, etc.

**What is new in this thesis?**

This thesis provides several contributions to chemoinformatics and and the design and discovery of potential antimalarial drugs:

1. **A database of antimalarial compounds** has been collected and rigorously curated with the help of our colleagues Dr. Elisabeth Davioud-Charvet (Strasbourg, France) and Dr. Birgit Viira (Tartu, Estonia). The collected set contains more than 15000 molecules, both recently synthesized and characterized in the Laboratory of Bioorganic and Medicinal Chemistry (University of Strasbourg) and coming from public ChEMBL database. All compounds have recorded experimental antimalarial activity value.

2. **Annotation by modes of action**. The mechanistic information on antimalarial compounds is sparse. Through careful bibliographical research, 1140 of the molecules in the collected database have been annotated by a hypothetical or an experimentally confirmed mechanism of action.

3. **Methodology of universal GTM**. The particularity of GTM algorithm has allowed us to develop a strategy of optimization of a GTM manifold that is able to simultaneously model several related or unrelated properties. This has led to the creation of "universal" maps: maps that are polypharmacologically competent with regard to more than 500 biological activities.

4. **Analysis of chemical space of antimalarial compounds**. The above mentioned strategy has been applied in order to create maps of antimalarial chemical space, which are able to effectively separate molecules active against *Plasmodium* from the inactive ones, as well as the compounds corresponding to different mechanisms of action. The maps allow to delineate zones corresponding to each target, and have helped us to correct erroneous annotations found in the database.

5. **Predictive models for redox potential of naphthoquinones**. Naphthoquinones are a family of antimalarial compounds that act *via* the oxidative stress in the parasite. With our colleagues from the Laboratory of Bioorganic and Medicinal Chemistry (University of Strasbourg, France) we have compiled a set of quinones and related molecules and have built the predictive models for their redox potential.

6. **Chemical space visualization software**. We have implemented a prototype web-based visualization tool that may be used to illustrate the obtained maps, the distribution of molecules on them, and allows to interactively navigate through the map.

All these developments and results are discussed in detail in the dedicated chapters. Before delving into the technicalities, let us begin by describing the primary focus of our research – malaria, and all that is known and done up to this day in the field of antimalarial research.

# Chapter 2

# Malaria and antimalarial drugs

Malaria remains one of the most severe infectious diseases that disproportionately affects the public health and economic welfare of the world's poorest communities (cf. Figure 2.1). It causes symptoms that typically include fever, feeling tired, vomiting, and headaches. In severe cases it can cause yellow skin, seizures, coma, or death. The causative agents of malaria belong to five species of protozoan parasites of the genus *Plasmodium* [2], namely *Plasmodium falciparum*, *Plasmodium vivax*, *Plasmodium malariae*, *Plasmodium ovale*, and *Plasmodium knowlesi*. The most prominent and dangerous killer among them, *P. falciparum*, is responsible for most severe forms of the disease, such as

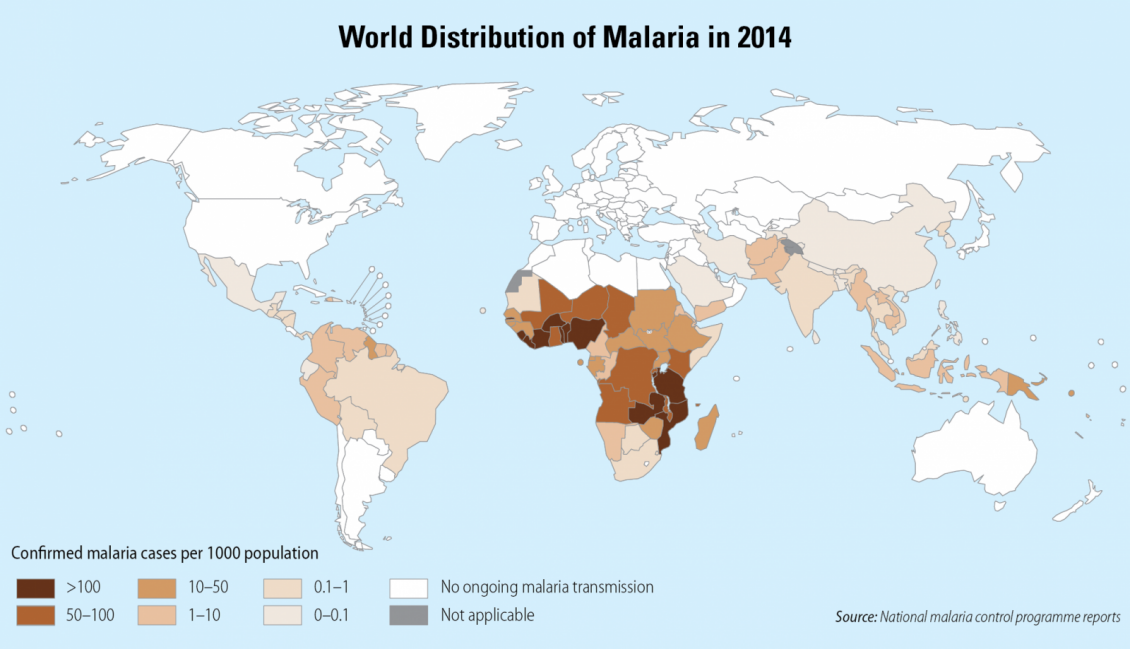



FIGURE 2.1: World distribution of malaria in 2014. According to the World Malaria Report 2014 [1], 97 countries and territories around the world are affected by malaria transmission.

cerebral malaria. Also, *P. vivax* malaria represents an important public health challenge because it has a wider geographical distribution than *P. falciparum* and the particularities of the parasite's life cycle lead to a more complicated relapsing form of malaria. Most malaria cases in 2015 are estimated to have occurred in the WHO African Region (88%), followed by the WHO South-East Asia Region (10%) and the WHO Eastern Mediterranean Region (2%) [3]. The disease is most commonly transmitted by an infected female *Anopheles* mosquito. Previously limited to tropical countries, in particular to Sub-Saharan Africa, the disease is now progressing in non-endemic regions [4] due to both global warming and migration flows. According to the WHO, 214 million clinical cases of malaria have been reported worldwide in 2015 giving rise, to 438,000 deaths per year that represents a 48% decline in mortality since 2000.

In 2015, whereas malaria is still the fourth highest cause of death, accounting for 10% of child deaths in sub-Saharan Africa, an equally impressive drop in malaria-related morbidity has been observed [3]. These spectacular statistics are partly due to the massive use of insecticide-treated bednets (ITNs) and artemisinin-based combination therapies (ACTs). In sub-Saharan Africa, of the million cases averted due to malaria control interventions it is estimated that 69% were accounted for the use of ITNs, 21% for ACT and 10% for indoor residual spraying. Unfortunately, decreases in case incidence and mortality rates were slowest in countries that had the largest numbers of malaria cases and deaths in 2000. Reductions in incidence need to be greatly accelerated in these countries if global progress for malaria elimination is to improve [3]. Furthermore, since 2008 there has been worrying decrease in the efficacy of ACT treatment in South-East Asia due to the emergence of drug-resistant parasites, which endangers the recent achievements and threatens the world's malaria control and elimination efforts.

While eliminating the asexual stages of plasmodial infections is the focus of treatment of individual symptomatic patients, at a population level, limiting the transmission of malaria, and in particular, the transmission of resistant parasites is pivotal for decreasing the community's burden of malaria [2]. In 2015 the Global technical strategy for malaria 2016–2030 was adopted by the World Health Assembly to target reductions in malaria cases and deaths for approaching malaria eradication. To reach this ambitious challenge in areas affected by ACT and multidrug resistance, the strategy prioritizes the rapid interruption of transmission of parasites to mosquitoes by new drugs targeting essential metabolic pathways in the insect stages. In considering antimalarial drug effects on transmissibility, three different components need to be considered (a) activity against asexual stages and early gametocytes, (b) activity against mature infectious gametocytes and (c) sporontocidal effects in the mosquito [5]. This is of critical importance, given the increased fitness of ACT resistant parasites to persist for days to weeks

in a non-replicating state and to survive drug exposure long enough before recovery and transmission [6].

There are numerous drugs that were or are currently used to treat malaria, but all of them have their drawbacks. The licensed 8-aminoquinoline drug primaquine [7] and the blue dye – methylene blue [8] – known as the first synthetic antimalarial molecule discovered in the early 20th century, possess transmission-blocking properties, but these drugs, in addition to other liabilities, produce either hemolytic anemia in individuals with glucose-6-phosphate deficiency or unsatisfying *in vivo* efficacy when used alone in humans, respectively. Chloroquine, the popular, licensed and widely used antimalarial drug in the first half of 20th century, had known restricted use because of the significant loss of efficacy in resistant parasite strains worldwide [9]. A plethora of 4-aminoquinoline analogues and derivatives, and many other classes of compounds have been developed since, and studied to overcome the drug resistance. For many existing antimalarial drugs that currently work well against malaria, the precise biological targets that they affect are often not entirely known. The WHO now recommends the use of ACTs, combining artemisinin drugs series with compounds of other mode of action, to ensure the effectiveness of the therapy, however, due to inadequate drug exposure [10] the resistance towards combination drugs has been recorded [11].

Vaccine remains the most efficient strategy for the eradication of infectious diseases, as was the case with smallpox and soon poliomyelitis. By their very nature, protozoa are more complex organisms than bacteria and viruses, with more complicated structures and life cycles. This presents problems in vaccine development but also increases the number of potential targets for a vaccine. Many malaria vaccine candidates have been tested or have been under development since the 1940s [12]. The only approved vaccine as of 2015 is RTS,S [13], developed by PATH Malaria Vaccine Initiative (MVI) and GlaxoSmithKline (GSK) with support from the Bill and Melinda Gates Foundation. It, nevertheless, has a relatively low efficacy (26–50%) [14]. PfSPZ vaccine [15] is a candidate malaria vaccine developed by Sanaria using radiation-attenuated sporozoites to illicit an immune response. The PfSPZ vaccine candidate has granted fast track designation by the U.S. Food and Drug Administration in September 2016 [16].

Since the complete genome sequencing of *P. falciparum* in 2002, partial genome sequencing of other *Plasmodium* species (*P. yoelii*, *P. vivax* and *P. knowlesi*) and various species races of *Anopheles gambiae*, the mosquito that transmits *P. falciparum* infection, malaria research has reached the post-genomic era [17]. Identification and elucidation of drug targets from the genome might be essential to enable an acceleration of preclinical candidates into the drug and vaccine pipelines. Genome sequences allowed increased

knowledge of the basic biology of malaria parasites, and opened new avenues of research in chemobiology to lead to new, rational therapeutic approaches. Novel genetic approaches [18] associated to phenotypic and chemoinformatic [19] screening can be used to discover new targets quickly and successfully, with the ultimate goal of finding new antimalarial chemical entities.

Now let us deliberate on the life cycle of the *Plasmodium* parasite and known and used antimalarial agents to give an overview on the state-of-the-art of the antimalarial drug design.

## 2.1 *Plasmodium* life cycle

During its development, *Plasmodium* goes through several stages to ensure the development of individuals and the sustainability of the species. This section describes the multiplication stages that occur in humans and in mosquitoes [2]. Figure 2.2 shows schematically the parasite's life cycle.

In human, two phases of asexual reproduction are distinguished: hepatic stage and blood stage.

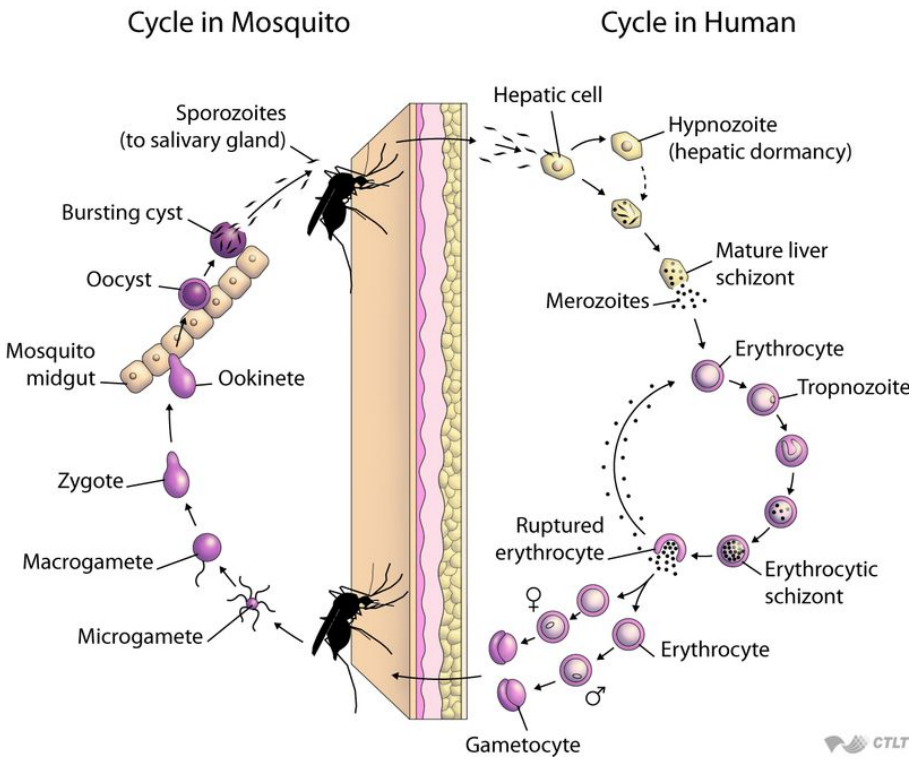



FIGURE 2.2: Schematic representation of *Plasmodium* parasite's life cycle. Human stages (hepatic and asexual blood stage) are on the right, the sexual stage in mosquito (sporogeny) is on the left. [20]

The first stage is essential and takes place in the liver. *Anophele* mosquito's female inoculates through a bite ***sporozoites*** living in its salivary gland. Through the bloodstream they are transported into the liver to multiply, unless they are destroyed by the immune system. Hepatic cells is the only possible place for sporozoite development. The duration of this intrahepatic multiplication is 7 to 15 days depending on the *Plasmodium* species. The product of this multiplication is a ***schizont*** containing thousands of ***merozoites***. The parasite escapes the immune detection by surrounding the merozoites in small bags formed by the host's liver membrane cell called ***merosome***. These then enter the bloodstream and are thus able to infect red blood cells. The hepatic stage is clinically silent and produces no symptoms.

The life cycle of *Plasmodium vivax*, a common malaria species found in India, Southeast Asia and South America, and *Plasmodium ovale* species is peculiar, because their schizonts are able to enter dormant state in the liver for months or even years. They are called ***hypnozoites***.

Once the host's red blood cells (RBC) are infected, the second phase of multiplication – asexual blood stage or ***erythrocytic schizogony*** – starts. Multiplication in RBCs lasts several days, depending on the parasite species. It results in the lysis of many red blood cells and causes the symptoms of malaria: anemia, fever, headache, vomiting, convulsions and even coma. During several multiplication cycles in blood, merozoites develop into ***gametocytes***, the sexual stages of the parasite.

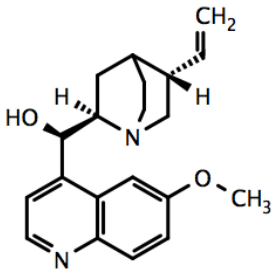
By taking a blood meal, the female mosquito ingests blood containing gametocytes. All other blood's components are digested in the stomach of the mosquito, except the gametocytes. In the mosquito gut gametocytes develop into gametes and these then form a zygote. Fertilization will transform the zygote into a motile ***ookinete***, which penetrate the mosquito's stomach wall and takes the ***oocyst*** form. Thousands of sporozoites then form within the oocyst, which will migrate actively in the salivary glands and become infective. This stage is referred to as ***sporogeny***.

## 2.2 Modes of action of antimalarial drugs

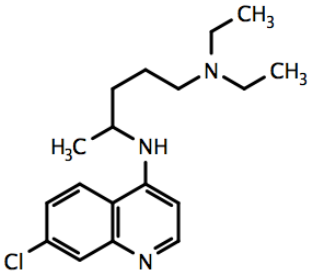
Antimalarial drugs are used for the treatment and prevention of malaria infection. Most antimalarial drugs target the erythrocytic stage of malaria infection, which is the phase of infection that causes symptomatic illness. It is practical to consider antimalarials by chemical structure since this is associated with important properties of each drug, such as mechanism of action. This section gives an overview of widely used antimalarial drugs and their modes of action.

### 2.2.1 4-aminoquinolines and related arylaminoalcohols

One of the oldest classes of antimalarial agents is the 4-aminoquinoline derivatives. The first compound – a natural alkaloid **quinine** – was extracted from the bark of *cinchona* tree (native to the Andean region of South America) in the late 1600s [21]. It has been used for treatment of fevers caused by malaria or other illnesses long before the identification of malaria parasite in 1880s [22]. However, it was impossible to mass-produce quinine due to its complex chemical structure: the first full stereochemical synthesis with sufficient yield was carried out in 2001 by Stork [23]. Therefore, compound with similar properties were always in need.




Quinine




Chloroquine

The first synthetic analogue of quinine – **chloroquine** – was synthesized in 1934 [9]. It has become the most widely used antimalarial drugs for some decades [24] due to its high efficacy, limited toxicity and simple and efficient synthesis. However, due to the development and spread of parasite resistance to this drug [25], it has seen the fall in its value in recent years.

Despite chloroquine and analogues being used for the treatment of malaria for a long time, their mechanism of action is not yet fully understood [26]. Chloroquine is only active against the erythrocytic stage of the parasite [27]. Various modes of action have been proposed for the drug: inhibition of protein synthesis [28], food vacuole phospholipases [29], aspartic proteinases [30], *etc*. Since the 1950s the 4-aminoquinolines are considered to inhibit the hemozoin formation in the food vacuole [31, 32]. *Plasmodium* parasite, in order to obtain essential aminoacids, degrades hemoglobin in the host's red blood cell. Heme generated in the process is toxic for the parasite. Several different pathways of heme detoxification (see Figure 2.3 for schematic representation) have been proposed [33]:

- sequestration of the free heme into hemozoin, or the malarial pigment;
- a degradation facilitated by hydrogen peroxide within the food vacuole;

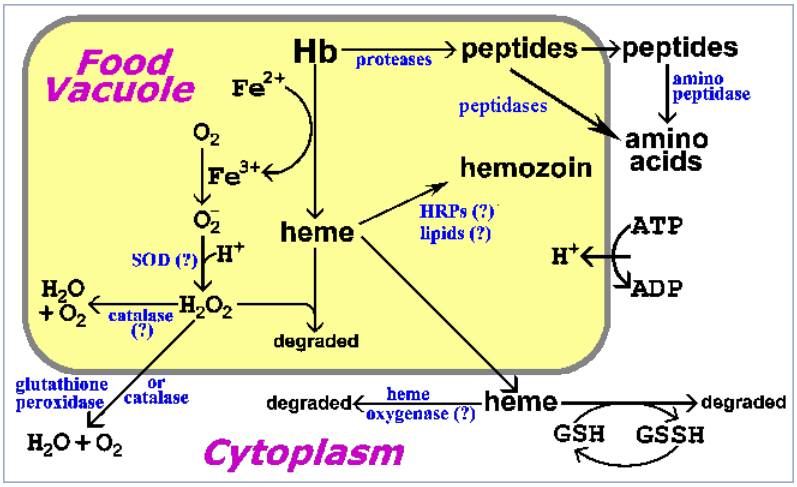



FIGURE 2.3: Pathways of heme detoxification in the food vacuole of *Plasmodium* parasite. Possible proteins implicated in the process are indicated with ? [35].

- a glutathione-dependent degradation which occurs in the parasite's food vacuole [34];
- and possibly, a heme oxygenase which has been found in *P. berghei* (rodent parasite) and *P. knowlesi* (simian parasite), but not in *P. falciparum*.

Hemozoin formation is a biocrystallization process. The first step of this is the formation of a dimeric β-hematin, which then seeds the crystallization process. The structure of hemozoin is represented on Figure 2.4

It has been shown that aminoquinolines bind to heme reversibly through $\pi - \pi$ interactions [38]. This interrupts the chain extension of the polymer, blocking further sequestration and detoxification of heme (Figure 2.5).

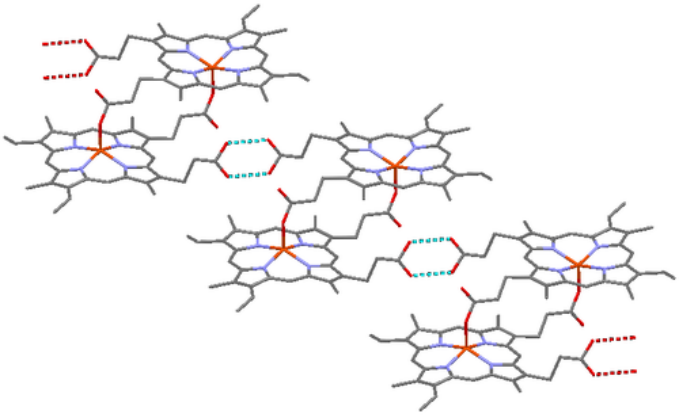


FIGURE 2.4: Structure of hemozoin – the heme polymer, the major detoxification product of the parasite's feeding process [36, 37].

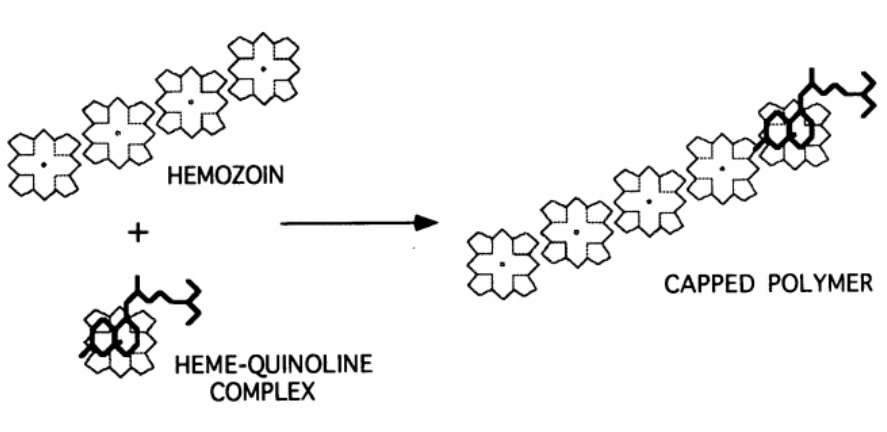
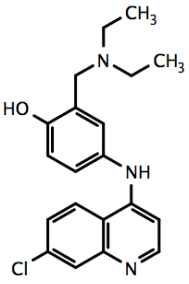
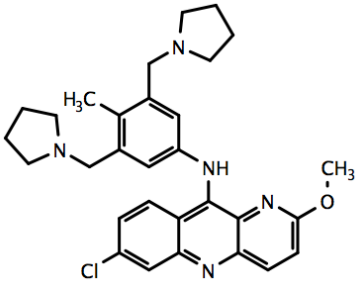


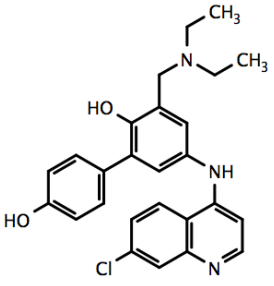

FIGURE 2.5: Supposed mechanism of action of aminoquinolines. The drug binds to the hemozoin through $\pi - \pi$ interactions, preventing further growth of the polymer [39].



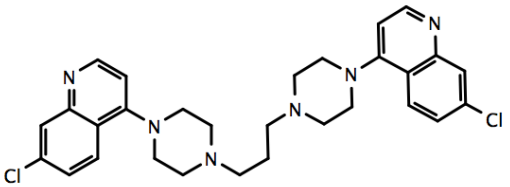

Amodiaquine




Pyronaridine



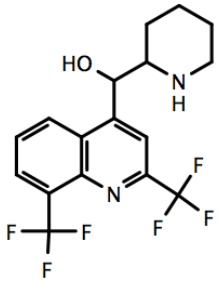

Tebuquine



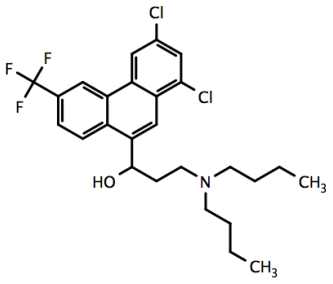

Piperaquine

The resistance of the parasite to chloroquine has been reported in 1950s. As with the mode of action, the mechanism of drug resistance is still open to discussion. The most renowned hypotheses for the resistance development are due to specific proteins: *P. falciparum* chloroquine resistance transporter (PfCRT) [40, 41] and *P. falciparum* multi-drug resistance homologue (PfMDR1) [42]. To overcome this resistance, novel aminoquinoline drugs have been developed since: **amodiaquine**, **piperaquine**, **tebuquine**, **pyronaridine**, and others. These analogues, however, have their drawbacks. Amodiaquine, a phenyl substituted analogue, has restricted clinical use due to hepatotoxicity [43], and has a cross-resistance with chloroquine [44]. Tebuquine series, while being highly active, have long half-lives, and have shown unacceptable toxicity profiles [45].

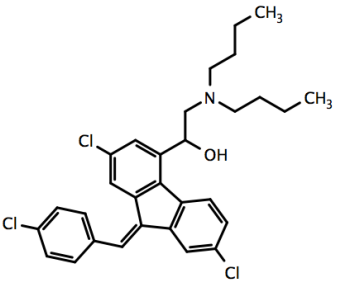
The most prominent drugs that correspond to the broad family of arylaminoalcohols are **mefloquine**, **halofantrine** and **lumefantrine**. These drugs are believed to share the mechanism of action with 4-aminoquinolines – the interaction with heme and inhibition of hemozoin formation and heme detoxification [46]. These compounds have seen the emergence of parasite resistance, both individual and cross-resistance within the family or to related families, and are thus not used for monotherapy, but as a part of combination therapies [47, 48]. However, mefloquine has unacceptable psychotic side-effects and therefore has been recently called to be discarded from the therapeutic market.




Mefloquine




Halofantrine




Lumefantrine

### 2.2.2 Antifolates

Antifolates is the general term used for drugs that block the synthesis or conversion of folic acid and its derivatives. Folate derivatives are important cofactors for the production of deoxythymidylate (dTMP) and, therefore, the synthesis of DNA [49]. Currently, disruption of folate metabolism is central in anticancer therapy [50], since rapidly dividing cells rely on the folates availability.

*Plasmodium falciparum* parasite is unable of getting pyrimidine from outside and thus relies completely on the *de novo* synthesis of dTMP. Therefore, the folate pathway is essential for parasite's survival [51]. The two most important for the folate biosynthesis enzymes are dihydrofolate reductase (DHFR) and dihydropteroate syntase (DHPS).

The archetypal DHFR inhibitors are **proguanil** and **pyrimethamine**. Proguanil was the first discovered antimalarial antifolate (1940s) [52]. It is a pro-drug, and is metabolized to its active triazine form, **cycloguanil** [53]. It was first used as a prophylactic agent, but, due to mutations in DHFR and decrease in its activity, it is not used as a monotherapeutic drug [54]. Pyrimethamine was initially synthesized as an anticancer drug, but was then identified as an antimalarial due to its structural similarity to proguanil [55]. Chlorproguanil, a chlorinated derivative of proguanil, generates through *in vivo* metabolism chlorcycloguanil, which is more potent than both cycloguanil and pyrimethamine. These drugs were especially effectively used in combination with the inhibitors of DHPS.

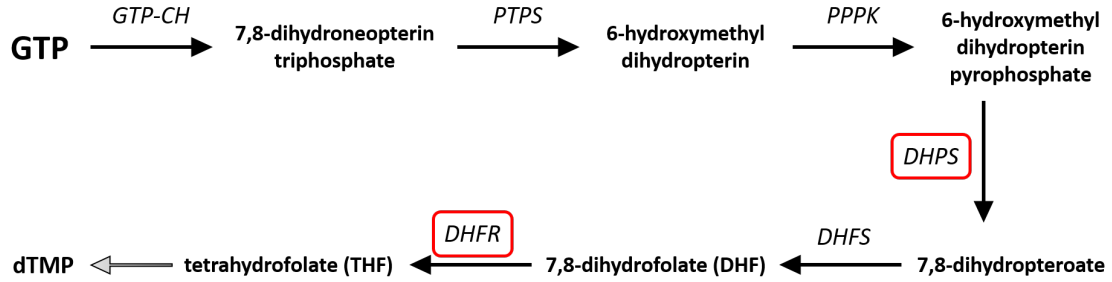
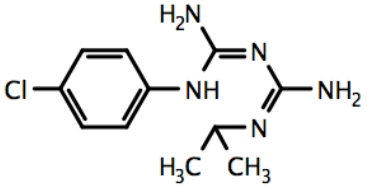
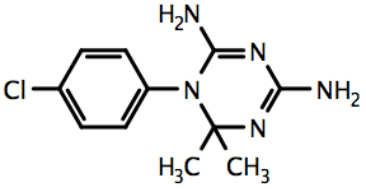


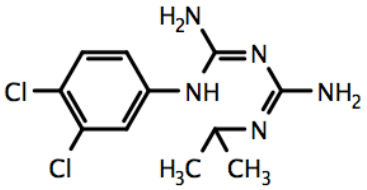

FIGURE 2.6: The *de novo* folate biosynthesis pathway of *P. falciparum*. Guanosine triphosphate (GTP) transforms into 7,8-dihydroneopterin triphospate by hydrolysis with GTP cyclohydrolase (GTP-CH), then via a cascade of tansformations by pyruvolytetrahydropterin synthase (PTPS), hydroxymethyldihydropterin pyrophosphokinase (PPPK), and dihydropteroate synthase (DHPS) forms dihydropteroate. Eventually, dihydrofolate syntase (DHFS) and dihydrofolate reductase finish the chain and give folate derivatives. Important drug targets are highlighted.



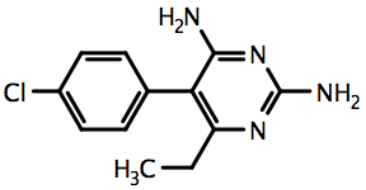
Proguanil



Cycloguanil

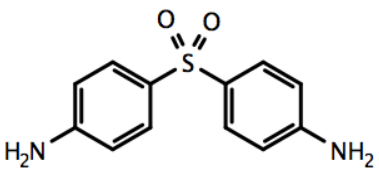


Chlorproguanil

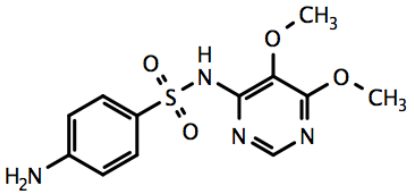


Pyrimethamine

DHPS inhibitors are sulpha-based and contain sulphonamide or sulfone groups [56]. The most important are **sulfadoxine** and **dapsone**. Attempts were made to use the DHPS inhibitors alone for the treatment of malaria [57], but due to low efficacy and high toxicity, these attempts were abandoned. Mostly they are used in combination with DHFR inhibitors in view of their synergy [58].



Dapsone



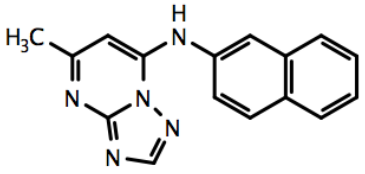
Sulfadoxine

While the emergence of antifolate-resistant strains of *P.falciparum* parasite has also been recorded [59], the folate pathway opens many opportunities for drug discovery. The activity of antifolate may be enhanced through combination or chemosensitization [60]. Some anticancer antifolates may also be used to treat malaria [61, 62].

### 2.2.3 Drugs affecting the electron transport chain

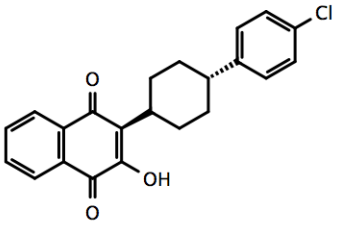
One of the interesting and selective targets for anti-plasmodial action is mitochondrial electron transport chain (ETC). Mitochondria in parasites are divergent from their mammalian counterparts. The main mobile electron carrier for ETC in malaria parasites is ubiquinone (coenzyme Q). Coenzyme Q is required as the electron acceptor by five enzymes: type II NADH dehydrogenase (NDH2), dihydroorotate dehydrogenase (DHODH), glycerol-3-phosphate dehydrogenase (GPDH), succinate dehydrogenase (SDH), and malate-quinone oxidoreductase (MQO). The reduced ubihydroquinone is oxidized by the cytochrome $bc_1$ complex, which is an important step for provision of coenzyme Q the abovementioned enzymes.

Out of these proteins, two are considered the most essential for drug discovery: cytochrome $bc_1$ complex [63] and DHODH [64]. DHODH participates in the pyrimidine biosynthesis pathway of the parasite, and since *P. falciparum* is unable to salvage pyrimidine, this is essential for its life cycle. Recent studies have revealed that different *P. falciparum* strains have varying dependency on the ETC, and thus certain ETC inhibitors have lower or no potency. It has been established that parasite-specific DHODH inhibitors such as triazolopyrimidines (for example, DSM1) work well as selective agents and are applicable to all parasite strains [65, 66].

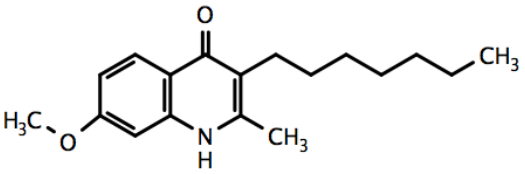



Triazolopyrimidine (DSM1)

The most prominent inhibitor of cytochrome $bc_1$ complex is a naphthoquinone drug – **atovaquone** [63]. Other compounds of similar structure (e.g. quinolone **endochin**) have also been used as antimalarial agents. However, atovaquone-resistant parasites have soon emerged [67], and atovaquone has since been used in combination with other drugs, for example, proguanil, the combination registered as Malarone [68]. Cytochrome *b* mutations and high cost of Malarone has, nevertheless, led to the restriction of its use to a prophylactic drug.




Atovaquone

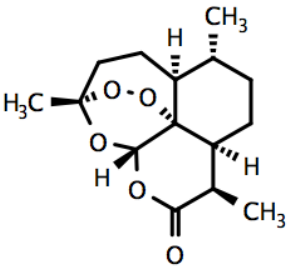



Endochin

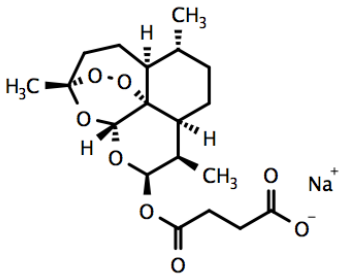
### 2.2.4 Artemisinins

Artemisinin comes from the plant, *Artemisia annua* or *quinghao* in Chinese, and has been used in traditional Chinese herbal medicines for treating relapsing fever for almost two thousand years [69]. The rediscovery of antimalarial properties of artemisinin in 1970s has made it the most useful drug for most malarial illnesses [70]. Prof. Youyou Tu, who has discovered the method for extraction of active ingredient from the plan, isolated it for structural analysis and discovered more active dihydroartemisinin, has been awarded the Nobel Prize in Physiology or Medicine in 2015 [71].

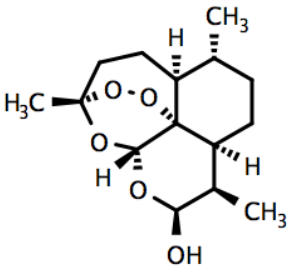
Now many semisynthetic derivatives are available, and the artemisinin derivative-based combination therapy (ACT) is now used as treatment for both uncomplicated and severe malaria in almost all of the countries with endemic *P. falciparum* [70]. The most prominent are **artemisinin**, **dihydroartemisinin**, **artemether**, **arteether**, **artesunate**.



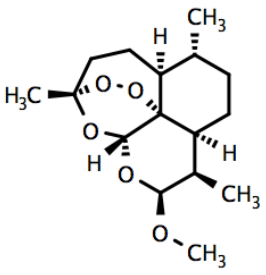

Artemisinin



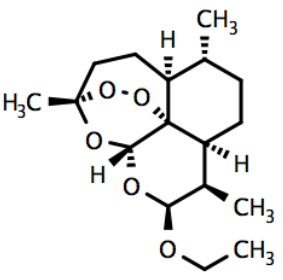

Artesunate




Dihydroartemisinin




Artemether




Arteether

The mechanism of action of artemisinin and its derivatives is still controversial and unclear [72–74]. Currently several concept are under debates, but they are not mutually exclusive, and it is probable that the anti-plasmodial activity of artemisinins is due to a number of factors at a time [75]. The most studied theories are the following:

- Heme pathway. As for the mode of action of aminoquinolines described above, artemisinins were shown to interact with the intracellular heme and inhibit the formation of hemozoin pigment both *in vitro* and *in vivo* [76]. To support this suggestion, the formation of heme-drug adducts has been demonstrated [77, 78]. However, the inhibition of heme crystallization has not been confirmed by other *in vitro* experiments [79, 80].
- Another theory is the alkylation and, thus, alteration of function of several key proteins in parasite, such as membrane transporters, proteases and other enzymes [81].
- Recent studies suggest PfATP6 enzyme as the biological target of artemisinins [82]. Nevertheless, *in vitro* studies with active synthetic analogues of artemisinin do no support this hypothesis [83].
- Studies of yeast have shown the effect of artemisinins on the mitochondrial electron transport chain [84–86]. These results are not supported by the studies of transgenic parasites [87].

Poor pharmacokinetic properties of artemisinins, particularly their short half-lives, translate into a substantial failure rates when used as monotherapy [88]. Therefore, it has been suggested to combine them with longer half-life partner drugs. In the recent decades, artemisinin and its derivatives have become the basis for antimalarial combination therapies [89, 90]. The WHO has defined antimalarial combination therapy as "the simultaneous use of two or more blood schizontocidal drugs with independent modes of action and thus unrelated biochemical targets in the parasite. The concept is based on the potential of two or more simultaneously administered schizontocidal drugs with independent modes of action to improve therapeutic efficacy and also to delay the development of resistance to the individual components of the combination" [91]. The WHO currently recommends five ACTs: artemether-lumefantrine, artesunate-amodiaquine, artesunate-mefloquine, artesunate-sulfadoxine-pyrimethamine, dihydroartemisinin-piperaquine [92]. However, the cost of ACTs is higher than chloroquine.

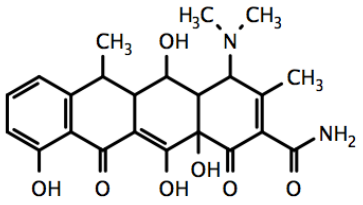
While the short half-life of artemisinins suggest that they would not be prone to the parasite developing resistance to it [93–95], cases of artemisinin resistance have been documented since 2006 in South-East Asia [96]. In addition, usage of ACTs, in theory,

should overcome the resistance by acting on several pathways at the same time, but the resistance to most commonly used partner drugs has been reported [97, 98]. Also, cross-resistance to drugs with related chemical structures is a concern. Therefore, the potentially severe implications of resistance to a drug to which there is currently no real alternative calls for cost-effective strategies to extend the useful life spans of currently available antimalarial drugs while at the same time investing into major efforts to develop novel compounds as a replacement for the artemisinins.

### 2.2.5 Various antibiotics

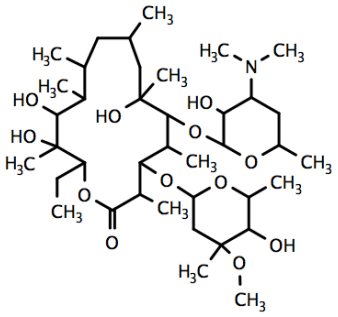
In addition to *Plasmodium*-specific drugs, a range of antibiotics have been shown to be efficient to treat malaria. In general, they target pathways within the apicoplast in apicomplexan parasites. They are usually not highly active, but are clinically practical is used in combination with other antimalarial agents. Additionally, unlike most specific drugs, they are safe to use for treatment of pregnant women an small children.

Tetracyclines have wide antimicrobial properties. They inhibit the protein synthesis by binding to the 30S ribosomal subunit. **Doxycycline** is an example of tetracycline antibiotic family that has demonstrated antimalarial activity [99]. It is relatively slow-acting, and is mostly used as a prophylaxis agent, however, when used in combination treatment, it may enhance the cure rates of conventional drugs. It is not recommended in pregnancy and children of age 8 years or less.




Doxycycline

Another example of a general antibiotic that has been found to be active against malaria is **azithromycin**, a semi-synthetic derivative of erythromycin [100]. It can, on the other hand, be safely administered by younger children or pregnant women. However, its high cost limits its use.



Azithromycin

## 2.3 *In silico* design of antimalarial compounds

*In silico* methods help to accelerate and facilitate the discovery and design of novel active molecules. However, such investigations demand sufficient amounts of high-quality experimental data, and in order to be exhaustive they should include as many chemotypes as possible.

In the field of antimalarial research, the problem of high-quality data is especially difficult. Currently, most studies available in the literature report quantitative structure-activity relationships (QSAR) for antimalarial compounds of one family at a time in a consistent biological experiment: substituted 4-aminoquinolines [101–103], naphthoquinone derivatives [104], endochins[105], sulfonamides [106], urea derivatives[107], analogues of other used drugs [108, 109]. Different methods and molecular descriptors may be used: quantum chemistry level QSAR is applied quite extensively, from DFT [110] to lighwight CODESSA quantum descriptors [111]. 3D QSAR is a popular approach for this difficult medicinal chemistry problem [112–114]. Nevertheless, the computational limitations of both quantum chemical calculation and 3D descriptors do not allow to investigate large amount of data.

Docking and pharmacophore models [115] are also often applied for the design of novel compounds and the mechanistic investigation of known series. For example, these techniques have been used to establish the possible mechanism of action of artemisinins [116] and other drugs [117].

Some studies of large datasets are found more often recently. Zhang et al. [118] reported to have developed QSAR models on a dataset of 3000 compounds. These models allowed to conduct a virtual screening of ChemBidge database, which has led to identification of 25 confirmed hits and novel antimalarial scaffolds.

One of the largest to date investigation is reported by Gamo et al [119]. They have done a screening of a large collection – millions – of compounds (provided by GSK) to identify

active molecules. This study is more an overview of the chemical space of antimalarial compounds (i.e. the span of active chemotypes) rather than a QSAR investigation. They have identified promising scaffolds for further development of new families of antimalarials compounds, and proposed as well modes of action hypotheses for some of these families. However, this overview was limited to the studied collection and did not include several critical antimalarial drugs (such as artemisinin) for validation and verification. Nevertheless, this assay data has been used in several works for both clustering and prediction of biologically relevant antimalarial targets and SAR information (by genetic methods [120] and network graph analysis [121, 122]).

Most recently, another open-source project for antimalarial drug discovery and investigation – MalariaBox [123] – has been initiated. The project assembles 400 diverse drug-like molecules for which antimalarial activity and mechanism of action are studied experimentally. This set provides data of high confidence that was verified by many laboratories all over the world, and that will stimulate further investigations for both *in silico* and experimental drug design projects.

# Chapter 3

# QSAR methodology

**Quantitative structure–activity relationship (QSAR)** modeling is one of the major computational tools employed in many fields such as drug discovery and toxicology [124]. As the name suggests, the goal of such studies is to find a mathematical function that relates the chemical structure to the property (such as biological activity) of the compound. This requires encoding a molecule's structural information in numerical form: a vector of *molecular descriptors* is used to define the position of the structure in *chemical space*. These descriptors must contain all the important structural information (nature of atoms, connectivity, ionization state, pharmacophoric features *etc*.) and risk ending up as a vector of very high dimensionality ($N > 1000$). The relation between molecular descriptors and a modeled property (or biological activity) is given as an equation:

$$P^{estim}(\text{molecule}) = f(\text{descriptors}).$$

This chapter gives an overview of the traditional QSAR modeling procedure, the notions of molecular descriptors, chemical space, model quality, and the modeling method that has been used in this work – Generative Topographic Mapping.

## 3.1 Molecular descriptors

There are different ways to represent a chemical structure, and each of them provides various information on the molecule of interest, and, consequently, different types of molecular descriptors [125] that can be generated.

1. **1D molecular descriptors** are derived from chemical formula (e.g. $C_2H_4O_2$). Those are the simplest descriptors associated with atomic properties such as atom counts

or molecular weight. Since these cannot discriminate between constitutional isomers, they are of limited usefulness.

2. **2D molecular descriptors**. These descriptors are based on 2D representation of a molecule — a molecular graph. A molecule is drawn as an ensemble of atoms (vertices of a graph) and bonds (edges). The descriptors associated with a 2D formula include the information on connectivity or properties that can be calculated from this representation, such as LogS or LogP values.

3. **3D molecular descriptors** are associated to the 3D structure of a compound. They include, for example, the van der Waals volume and the density. Alternatively, given a molecule embedded into a 3D lattice grid, other types of descriptors may be calculated, such as molecular interaction fields, or pharmacophoric fingerprints.

In this thesis, ISIDA Substructural Molecular Fragments (ISIDA SMF) [126] are used. These are 2D descriptors that encode the molecular structure by a vector consisting of occurrences of fragments of different types. The fragments may be linear sequences, augmented atoms (central atoms and their environment) or triplets, encoding information on atom types and/or bonds. Additionally, atoms may be colored by supplementary information: pharmacophore type, $\log P$ increment, force-field atom types, *etc* [127].

Another type of descriptors is developed in this work. Electronic Effect descriptors (EED) [128] represent the effect of the chemical environment on the electronic properties of a given atom or group. These will be elaborated in the dedicated chapter.

## 3.2 Model building

When the molecular descriptors for the given set of compounds are generated, the QSAR model can be built. The dataset used for that stage is called ***training set***. To train a model, a machine learning [129] algorithm is applied, and, depending on the task at hand, different paradigms of modeling are used.

When the goal of the study is to predict numerical or categorical value from descriptor vector, for example, predicting biological activity of a molecule from its structure, it is called **supervised** learning. In that case, the modeled property is known for all compounds of the training set, and the output of the model is the predicted value - numerical in case of regression tasks, and categorical in case of classification. Methods such as Artificial Neural Network (ANN) [130], Multilinear Regression (MLR) [131], Random Forest (RF)[132] or Support Vector Machines (SVM) [133] are often used for supervised

model building paradigm. The training of the model here consists of fitting a function to minimize the error of prediction.

**Unsupervised** learning, on the other hand, is where one only has input descriptor vectors and no corresponding output variables. The goal for unsupervised learning is to model the underlying structure or distribution in the data in order to learn more about the data. Unlike supervised learning above, there is no correct answers and there is no teacher. Algorithms are left to their own devises to discover and present the interesting structure in the data. Mostly, the unsupervised learning methods are about clustering [134] the data to discover the inherent groupings in the dataset, with methods such as $k$-means [135].

Another subtype of unsupervised methods, very useful for describing and interpreting the data at hand, are the visualization methods [136]. As it is impossible for human to perceive the high dimensional space defined by the molecular descriptors, it is useful to use dimensionality reduction methods that will generate a smaller number of new features. For visualization purposes, the final number of dimensions should be 2 or 3. Dimensionality reduction methods may also be used for reducing the number or generating more general features for the modeling, which provides a way to avoid the dimensionality curse [137] as well as computational costs that may occur when working with high-dimensional data. There are many methods for dimensionality reduction and visualization: Principal Components Analysis (PCA) [138], Linear Discriminant Analysis (LDA) [139], Linear Multidimensional Scaling (MDS) [140], Locally Linear Embedding (LLE) [141], Self-Organizing Maps (SOM) [142], Generative Topographic Mapping (GTM) [143].

## 3.3 Model quality measures

Once a model is built, its quality and robustness should be evaluated by one or several statistical parameters [144] by applying it to a designated test set and comparing the predictions to the actual experimental values. These parameters are different for regression and classification models.

For regression models, a model's robustness would be assessed by the numerical deviation of predicted value from a known experimental. The most commonly used characteristic is the ***Root Mean Squared Error (RMSE)*** that represents the average error of prediction:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N} (y_{exp,i} - y_{pred,i})^2}{N}}. \tag{3.1}$$

Here, $N$ is the size of considered dataset, $y_{exp,i}$ and $y_{pred,i}$ are the experimental and predicted activity values for $i$-th molecule.

In addition, one may use the ***determination coefficient*** $R^2$, that demonstrates the correspondence between experimental and predicted values for the set. $R^2 = 1$ represents the ideal fit, where all values are predicted perfectly, and the lower its value, the worse are the predictions. It is given by the following formula:

$$R^2 = 1 - \frac{\sum_{i=1}^{N} (y_{exp,i} - y_{pred,i})^2}{\sum_{i=1}^{N} (y_{exp,i} - < y_{exp} >)^2}, \tag{3.2}$$

$< y_{exp} >$ denotes the mean activity value of the dataset.

In case of classification models, a model's quality is evaluated based on the number of entities with correctly predicted class. Most classification tasks are reduced to binary classification, where only two classes are present. In that case, to evaluate the model, one constructs the confusion matrix. It is a table where cells contain number of entities correctly ('True') or incorrectly ('False') predicted by the model to be either of the classes (annotated 'Positive' or 'Negative', without any correspondence to actual positive or negative effect in general):

| | | Prediction outcome | |
|---|---|---|---|
| | | Class 1 | Class 2 |
| Actual value | Class 1 | True positive ($TP$) | False negative ($FN$) |
| | Class 2 | False positive ($FP$) | True negative ($TN$) |

From these numbers, a numerical characteristic of a model's quality may be calculated. In this work, we only use the ***balanced accuracy (BA)***:

$$BA = \frac{1}{2}\left(\frac{TP}{TP + FN} + \frac{TN}{FP + TN}\right). \tag{3.3}$$

This measure is convenient for sets where one class is predominant over the other, since it takes the rate of correct predictions for both in equal proportions. In the case of zero-rules model, where all entities are predicted to be one class, $BA$ is equal to 0.5, therefore, as a rule of thumb, a model with $BA$ inferior to 0.5 is considered low-quality, and the closer it is to 1, the better is the model.

### 3.3.1 Cross-validation

In the ideal case, each model should be validated on a specially picked separate test set. However, a separate data set can and will have a different statistical distribution than that which was used in in training stage, so the validation will be biased. Therefore, a cross-validation model evaluation procedure [145] is applied. The goal of cross-validation is to define a dataset to "test" the model in the training phase (*i.e.*, the validation dataset), in order to recycle the training data, to limit problems like overfitting, and to give an insight on how the model will generalize to an independent dataset (*i.e.*, an unknown dataset, for instance from a real problem), *etc*.

The cross-validation procedure involves partitioning a sample of data into complementary subsets, performing the analysis on one subset (called the training set), and validating the analysis on the other subset (called the validation set or test set). To reduce variability, multiple rounds of cross-validation are performed using different partitions, and the validation results are averaged over the rounds.

In this work, ***k-fold cross-validation*** (see Figure 3.1) is used for model evaluation. This technique consists of dividing the initial data sample into $k$ parts (*folds*) and alternatively using them for training ($k-1$ parts) and testing (the last part) the model. Eventually, every molecule is predicted as a part of test set exactly once, and these prediction values are used for calculating cross-validated statistical parameters ($BA$, $RMSE$, and $R^2$).

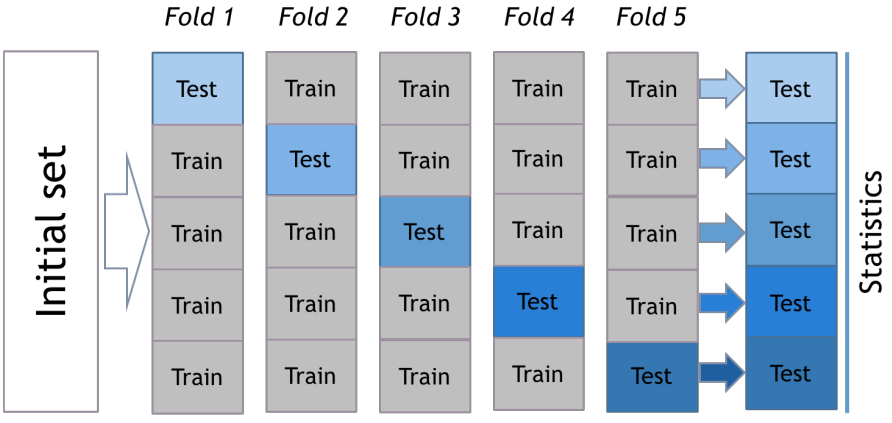



FIGURE 3.1: Schematic representation of 5-fold cross-validation procedure. Initial dataset is divided into 5 parts, on each fold a model is trained on 4 parts and is applied to the last one. Then, all predicted values are gathered for statistical evaluation.

## 3.4 Generative Topographic Mapping

The Generative Topographic Mapping (GTM) method is a probabilistic extension of Kohonen Self-Organizing Maps and is now a popular method of data visualization. The method was introduced by Bishop and Svensen [143, 146] in 1998. GTM models try to find a representation for the distribution of data in an initial $D$-dimensional data space $\boldsymbol{t} = (t_1, t_2, \ldots, t_D)$ by a number $L$ of latent variables $\boldsymbol{x} = (x_1, x_2, \ldots, x_L)$. To do this, a function $\boldsymbol{y}(\boldsymbol{x}; \mathbf{W})$ is applied to the points in the latent space and maps them into the data space (Fig. 3.2).

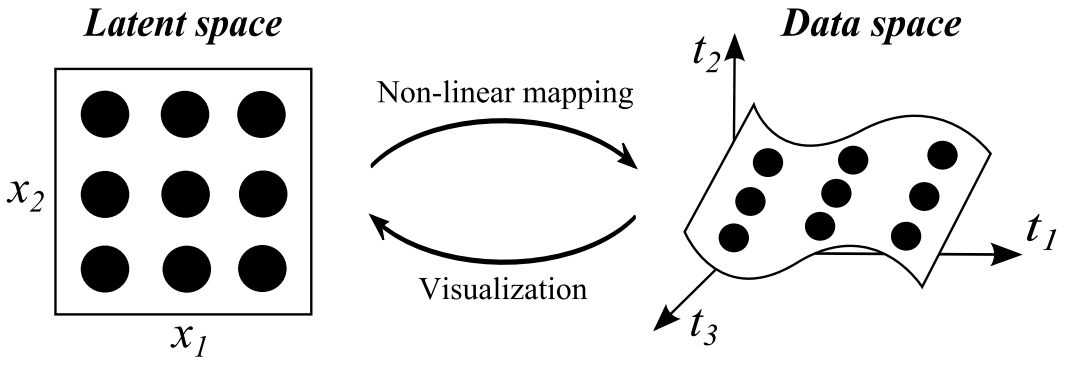



Figure 3.2: Schematic representation of the GTM algorithm. Here, latent space has 2 dimensions, data space has 3 dimensions.

In this thesis, GTM algorithm is applied through ISIDA/GTM software, implemented in FreePascal by H. Gaspar (version 2015) [147]. In this version, the non-linear mapping function is given as a grid of $M$ Gaussian activation functions (radial basis functions, or RBF), and the transformation is made as follows:

$$\boldsymbol{y_d}(\boldsymbol{x}; \mathbf{W}) = \sum_{m=1}^{M} W_{md} \exp\left(\frac{\|\boldsymbol{x} - \boldsymbol{x_m}\|^2}{2\sigma}\right) \tag{3.4}$$

where $d$ goes from 1 to $D$, $\mathbf{W}$ is the $M \times D$ weight matrix connecting RBF and data space points, $\boldsymbol{x_m}$ is the center of the $m$-th RBF, their number $M$ and width $\sigma$ are the parameters of the method.

For visualization purpose it is reasonable to take $L = 2$. In this case, the latent space may be viewed as a flexible "rubber sheet" (called *manifold*) embedded into the data space. Since the transformation function (3.4) is smooth and continuous, points which are close in the latent space remain neighbors in the data space. A given node in the 2D grid of the map is associated with the center of a normal distribution function with inverse variance $\beta$, that corresponds to the sampling of a random variable $\boldsymbol{t}$ with the following probability

density function:

$$p(\boldsymbol{t}|\mathbf{W},\beta) = \frac{1}{K}\sum_{k=1}^{K}\exp\left(-\frac{\beta}{2}\|\boldsymbol{t}-\boldsymbol{y}(\boldsymbol{x}_k,\mathbf{W})\|^2\right), \tag{3.5}$$

where $\boldsymbol{x}_k$ are the coordinates of $k$-th grid node in the latent space, $\boldsymbol{y}(\boldsymbol{x}_k,\mathbf{W})$ are the coordinates to which it was mapped in the data space, $\boldsymbol{t}$ spans data space and represents any data point.

The correspondence between GTM model and data is measured by *log likelihood* $\mathcal{L}$ — the logarithm of the probability with which the data could be generated from $\boldsymbol{t}$, which is a function of two adjustable parameters, $\mathbf{W}$ and $\beta$:

$$\mathcal{L}(\mathbf{W},\beta) = \sum_{n=1}^{N}\mathcal{L}_n = \sum_{n=1}^{N}\ln\left(\frac{1}{K}\sum_{k=1}^{K}\exp\left(-\frac{\beta}{2}\|\boldsymbol{t_n}-\boldsymbol{y}(\boldsymbol{x_k};\mathbf{W})\|^2\right)\right), \tag{3.6}$$

$\boldsymbol{t}_n$ is the position of $n$-th data point in the data space. The higher the value of $\mathcal{L}$, the better the manifold represents the data, so $\mathcal{L}$ is used as the maximized function for the training, which is achieved by the Expectation Minimization (EM) algorithm [148].

In GTM, a data point has a non-zero probability to be mapped into any node of the map. This probability is called the *responsibility* of $k'$-th node for $n$-th data point $\boldsymbol{t}_n$ and is calculated using Bayes' theorem:

$$R_{nk'} = p(x_{k'}|\boldsymbol{t}_n) = \frac{\exp\left(-\frac{\beta}{2}\|\boldsymbol{t_n}-\boldsymbol{y}(\boldsymbol{x}_{k'};\mathbf{W})\|^2\right)}{\sum_{k=1}^{K}\exp\left(-\frac{\beta}{2}\|\boldsymbol{t_n}-\boldsymbol{y}(\boldsymbol{x}_k;\mathbf{W})\|^2\right)} \tag{3.7}$$

Responsibility is normalized over the grid of nodes, so the sum of responsibilities for a given data point is equal to 1. This vector is used for both visualization and modeling purposes.

Classical GTM algorithm is however quite computationally costly when the number of molecules and descriptors is very high. To overcome this problem for the analysis of large datasets (hundreds of thousands compounds), an incremental variation of GTM algorithm has been introduced [149]. In this case, the initial dataset is divided into a number of blocks, and they are given to the algorithm one by one. In such a way, the manifold will train on one block at a time, consuming proportionally less memory. In the current work, incremental algorithm is used almost exclusively.

### 3.4.1 GTM-based regression and classification models

As GTM generates a vector of normalized responsibilities for each molecule, this vector can be turned into molecular descriptors. It will contain as many descriptors as there are nodes in the GTM grid and is used as a basis for designing the activity landscape. The landscape is obtained by adding the property of interest for the given training set as a third dimension or a "color" to the 2D map. After the landscape is computed, the GTM QSAR model may be built. There are different ways to develop a model, but all require the same procedure: building a manifold from the training set, "coloring" it by a property and then projecting the test set to calculate responsibilities and coordinates of test set molecules.

In **GTM-based regression models**[150] the predicted property value $\bar{A}_K$ for the node $K$ is calculated by using actual activities and responsibilities of the data set as follows:

$$\bar{A}_K = \frac{\sum_{i=1}^{N} A_i R_{iK}}{\sum_{i=1}^{N} R_{iK}}, \tag{3.8}$$

where $N$ is the number of molecules in the data set, $A_i$ is the experimental activity of $i$-th molecule, $R_{iK}$ is its responsibility in the node $K$ (see equation (3.7)). After the landscape is computed, the GTM QSAR model may be built. There are different ways to develop a model, but all require the same procedure: building GTM map from the training set, and then projecting the test set to calculate responsibilities and coordinates of test set molecules.

*Global activity landscape method* uses activity landscape and responsibilities of the test set for prediction. The predicted activity $\hat{A}_j$ is obtained by weighing all $K$ landscape activity values by responsibility of $j$-th molecule:

$$\hat{A}_j = \sum_K \bar{A}_K R_{jK}, \tag{3.9}$$

where $\bar{A}_K$ is the activity landscape value in $K$-th node, $R_{jK}$ is the responsibility of the molecule in that node.

In *local activity landscape method*, a kNN approach is applied to test set molecules. For each test molecule, $k$ nearest grid nodes are selected using Euclidean distance, and the predicted activity is computed as the average of activities $\bar{A}_k$ in these nodes:

$$\hat{A}_j = \sum_k \frac{\bar{A}_k}{k}. \tag{3.10}$$

Another approach is *kNN in 2D space*. For each test set molecule, the activity is calculated as the average over the activities $A$ of $k$ training set molecules, closest to it in the 2D latent space, so only the coordinates of all points is needed:

$$\hat{A}_j = \sum_k \frac{A_k}{k}. \tag{3.11}$$

**GTM-based classification models** [151] attribute a class to each node of the grid by averaging the responsibilities $R_{tk}(C_i)$ over $N_{C_i}$ training set compounds belonging to the $i$-th class in the latent space. This allows to compute the conditional probability $P(k|C_i)$ of finding a new instance close to a node $k$:

$$P(k|C_i) = \frac{\sum_t R_{tk}(C_i)}{N_{C_i}}. \tag{3.12}$$

One can then calculate the conditional probability $P(k|C_i)$ of class $C_i$ given node $k$ using the Bayesian theorem:

$$P(k|C_i) = \frac{P(k|C_i) \times P(C_i)}{\sum_j P(k|C_j) \times P(C_j)}, \tag{3.13}$$

where $p(C_i) = N_{C_i}/N_{tot}$ with $N_{tot}$ being the total number of compounds in the training set. These conditional probabilities are then used for the compound $q$ of the test set to estimate the $i$th class probability $P(C_i|q)$, by one of two approaches. One is Bayesian, where probabilities of each class are calculated as following:

$$P(C_i|q) = \sum_k P(C_i|k) \times R_{qk}, \tag{3.14}$$

then the class with the highest probability is assigned to the compound. The other approach is the $k$-nearest node, where simply the predominant class of the nearest node $k$ on the map is assigned to the compound $q$:

$$P(C_i|q) = P(k|C_i). \tag{3.15}$$

## 3.5 GTM visualization

Initially, GTM was only used as a visualization method [152, 153]. As it was shown in previous sections, now it is successfully used for QSAR modeling and dataset comparison, as well, but its advantage over other classical machine learning methods – the visualization aspect – is more than relevant.

The simplest form of GTM visualization is the distribution of molecules over the manifold. The mean positions of compounds are calculated from the responsibility vector as weighted average coordinates. They are then placed on the map as black dots (Figure 3.3). This kind of visualization provides information on map coverage, but is not very informative otherwise.

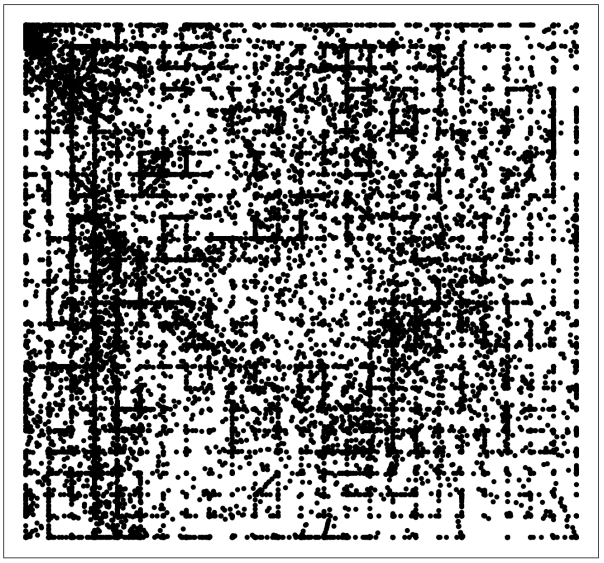


FIGURE 3.3: A simple visualization of a GTM. Black dots represent the mean position of compounds on the map. The compounds used and map construction process is explained in Section 5.1.

A more useful way of the map visualization is to construct ***activity landscapes***. As it was stated in the previous section, the nodes of the manifold may be *'colored'* by a property based on the responsibility distribution of the training set. Therefore, this can be visualized by adding a color or a third axis to the 2D representation of the manifold grid. Figures below show examples of activity landscapes for numerical (Figure 3.4) and categorical (Figure 3.5) properties. Additionally, the landscape is weighted by data density: the less molecules contribute to a specific node, the less intense the color is. A threshold may be given, so that nodes with lower density would define "blank zones" – nodes that bear no significant contribution to a property.

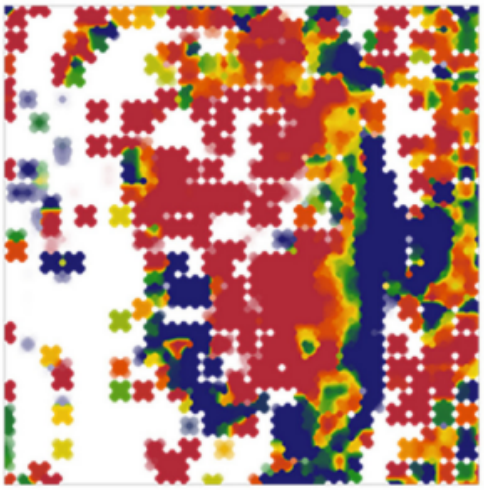


FIGURE 3.4: Example of a property landscape [154]. Here, a map for 2 million conformers is built and colored by energy, from −56 kcal/mol (in red) to −44 kcal/mol (in blue). Blank zones correspond to areas of low data density.

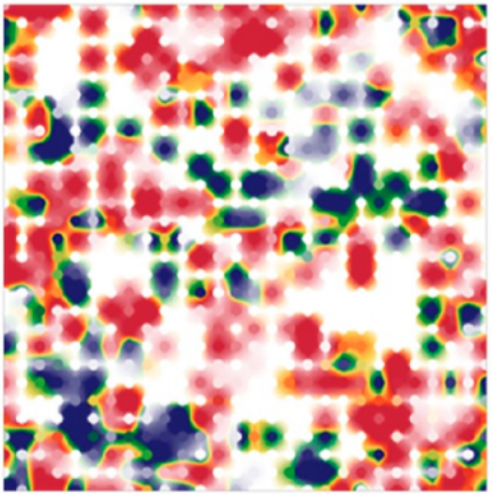


FIGURE 3.5: Example of a class landscape [155]. A map is built for antiviral compounds in ChEMBL, the colors represent the class – blue for compounds active against Lentivirus, red for inactives. Intermediate colors represent zones with mixed population.

Such density weighing represents a special type of applicability domain (AD) for GTM models. Applicability domain concept is important for QSAR modeling. By one of definitions, "the applicability domain of a (Q)SAR model is the response and chemical structure space in which the model makes predictions with a given reliability" [156]. In most cases it is understood that the application of a QSAR model should be a case of interpolation, not extrapolation, meaning that the compounds on which it is applied should be close in descriptor space to the molecules used in training stage. The "blank zones" of GTM represent exactly this: since they show areas to which too few molecules contribute, if test set molecules are projected there, they should be considered outside of AD for the map.

Other, more numerically elaborated kinds of AD also exist for GTM. One of them is especially useful for classification models, and was successfully used for BDDCS dataset [151]. The principle of this AD is that the nodes are considered within AD if one class has higher Bayesian probability than others. The threshold for this predominance is variable and defined by the user.

## 3.6 Chemical space analysis

As a visualization and modeling tool, GTM offers a possibility for not only quantitative prediction of a property, but also for a qualitative and insightful analysis of chemical

space [157]. It has been mentioned above how the visualization helps in identifying dense or sparse zones of chemical space which can be used for comparing datasets, highlighting missing or dominating chemotypes, etc. Alternatively, this analysis gives insight into the neighborhood behavior of the dataset, *i.e.* determining common scaffold or features for neighboring compounds of one or different classes, detecting SAR information-rich regions, and so on.

Visual inspection of the map is, however, impractical in case of large datasets. On the other hand, the mean position of a compound on the map is only representative in case of unimodal responsibility distribution, which is not always true. Additionally, responsibility vectors consist of real values, thus detecting molecules with strictly identical responsibilities is highly unlikely. Therefore, it has been proposed [155] to transform the vector into a discretized form, where responsibilities would be binned by intervals of 0.1, and assigned levels 0–10. Such discrete representation is called ***responsibility pattern (RP)*** of a molecule. Molecules with the same RPs are considered to belong to a same cluster or 'family', because they share some common structural motif which is the underlying reason for their mapping into the same map zone 'covered' by the RP. As already shown [155, 158], common structural motifs may range from precisely defined scaffolds or even specifically substituted scaffolds, to fuzzier ensembles of related, interchangeable scaffolds, to even fuzzier 'pharmacophore-like' patterns.

If an RP cluster appears to be enriched in active compounds, then the common structural motif defining the cluster is privileged with respect to that property. Quantitatively that can be measured as relative frequency of specific RP for given activity—class specificity ($CSP$):

$$CSP(RP, C) = \frac{f_C(RP)}{f(RP)} \tag{3.16}$$

where $f_C(RP)$ is the occurrence frequency of RP among compounds of class $C$, and $f(RP)$ is the pattern occurrence frequency within the set of compounds used to build the manifold (i.e. frame set). Otherwise, frequencies of RP in different classes or in different sets may be used for other types of analysis.

This analysis only makes sense for often encountered responsibility patterns. A threshold may be set for $CSP$ to define the most prominent privileged RPs, *i.e.* they are to be of interest if the threshold value for $CSP$ is achieved.

# Chapter 4

# Data collection and curation

Collecting an adequate database is the first and most crucial step of building a QSAR model. The search for antimalarial has seen an exponential rise in interest in the recent decades, especially with the addition of big pharmaceutical companies such as GSK or Novartis to the search of efficient antimalarial agents. Public ChEMBL database has compiled a separate set – ChEMBL Malaria – which assembles all available knowledge on the compounds and targets relevant to malaria. This knowledge base has over 4 million records of bioactivity for 282000 distinct compounds, measured in functional or single-protein (5980 targets in total) assays.

While the available data seems abundant, the main issue is its heterogeneity. As good practices of QSAR modeling [159] state, high heterogeneity of initial dataset compromises the model's integrity and thus should be avoided. Research in the malaria domain is still largely limited to the *in vivo* studies of small congeneric series of compounds, which complicates the data collection. Often, molecules are found active in one assay, but inactive in another with different conditions. An extreme care should be therefore taken in handling these data in QSAR studies.

Another important issue is the lack of knowledge of biological mechanism of action of the majority of antimalarial compounds. Most of the activity data reported in ChEMBL originate from functional or whole-cell assays, therefore not taking into account the biological target of a compound and only measuring its ability to kill the parasite. Furthermore, compounds may be occasionally annotated in target-specific assays, but most often these are results of independent experiments, not related to antimalarial activity. Out of 5980 recorded targets, only 114 concern *Apicomplexa* parasites, and only 90 are *Plasmodium*-related. 14 of them correspond however to phenotypic anti-*Plasmodium* assays and contain the majority of data (Figure 4.1).

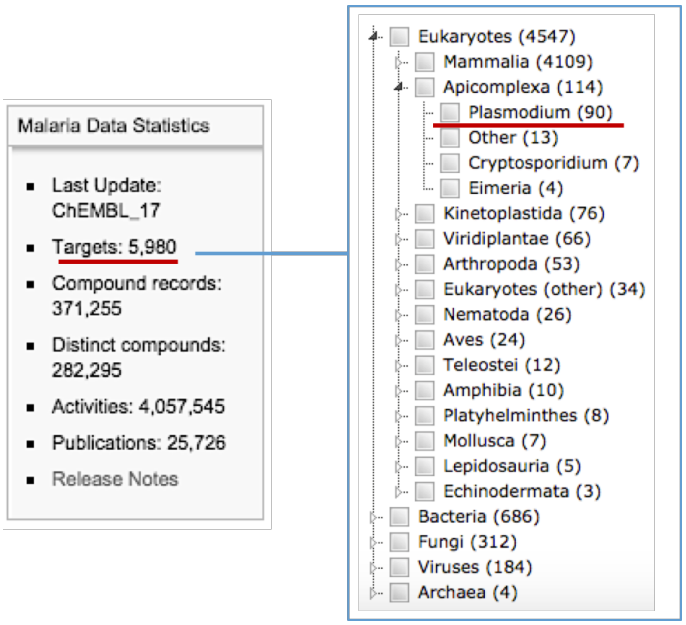



FIGURE 4.1: Malaria data statistics from dedicated ChEMBL page (`www.ebi.ac.uk/chembl/malaria/`). 4 million records are available for 280000 distinct compounds. However, among 5980 recorded targets only 90 concern *Plasmodium* parasite.

This chapter describes the process of the collection of a relevant dataset of antimalarial compounds, its curation and homogenization, the criteria of activity and biological target annotation.

## 4.1 MalariaDB and ActivityDB

Two primary sources of data are used in this work – experimental data from our collaborators and ChEMBL database [160]. A large database of high-confidence antimalarial compounds data is collected during the thesis, which is called henceforth **MalariaDB**. A subset of this database, **ActivityDB**, assembles data for compounds measured with high precision in similar experimental condition in an attempt to create a dataset for robust prediction of antimalarial activity. The composition of MalariaDB and ActivityDB is described in this section.

ChEMBL database has a dedicated section – ChEMBL Neglected Tropical Diseases (found on `www.ebi.ac.uk/chemblntd`) – that is a compilation of primary screening and medicinal chemistry data for endemic tropical diseases, especially parasites such as *P. falciparum*, *L. donovani*, *T. brucei* and *T. cruzi*. Two datasets from this repository are extracted into MalariaDB.

The first one is GlaxoSmithKline's Tres Cantos Antimalarial Set (TCAMS) [119]. The set comes from a screening study of over 2 millions compounds against *P. falciparum*

strains 3D7 and D2 in human erythrocytes. 13256 reported hits inhibit the growth of the parasite by more than 80% at 2 $\mu$m concentration. Since data comes from a single screening assay, the measured activities are considered as-is.

The second considered dataset is Medicine for Malaria Venture Pathogen box (MMV) [123]. The set includes 400 diverse compounds with confirmed biological activity against a number of parasitic diseases, including malaria. The assays were not entirely performed following one protocol.

These two sets were selected not only for the high quality of measured data, but especially for the target annotations of screened molecules. The annotation sources, curation and validation are discussed in the following section.

The last part of MalariaDB comes from the collaboration with Dr. Elisabeth Davioud-Charvet's team (Laboratory of Bioorganic and Medicinal Chemistry, University of Strasbourg, France) and Dr. Birgit Viira (University of Tartu, Estonia). The colleagues have provided us a set of over 200 molecules with antimalarial activity measured according to 11 different bioassays. In order to enrich the dataset, a search in ChEMBL database has been carried out for compounds with dose-response values obtained in similar conditions, which has led to addition of 86 sets with at least 50 entries. Due to the unreasonably high number of reported distinct activity measures, it has been decided to merge datasets if possible. Two main merging strategies were applied: 1) if the key experimental conditions and the nature of measured activity of two sets are the same, they are fused together; 2) if the reported activity values for at least three common compounds for two sets are the same, the sets are merged. The merging strategy is explained in detail elsewhere [161]. This resulted in 17 distinct datasets (Table 4.1) describing in total 2093 compounds (molecules may be shared between several sets). In each one, the entries are annotated 'active' if measured dose-response activity value is in submicromolar range, and 'inactive' otherwise. This part of the data is called henceforth ***ActivityDB***. The composition and some experimental details are given in Table 4.1 and in the Appendix.

## 4.2 Mode of action annotation

Out of 15462 molecules in the dataset, 1140 are annotated with a biological target or target family. This section explains the sources and confidence levels of annotations.

| Subset | Measured Property | *Plasmodium* Strain | Size |
|---|---|---|---|
| PS1 | pIC50 | 3D7 | 65 |
| PS2 | pIC50 | K1 | 126 |
| PS3 | pIC50 | Dd2 | 66 |
| PS4 | pIC50 | Dd2 | 70 |
| PS5 | pIC50 | K1 48 | 125 |
| PS6 | pIC50 | 3D7 | 120 |
| PS7 | pIC50 | 3D7 | 94 |
| PS8 | pIC50 | Dd2 | 143 |
| PS9 | pIC50 | K1 | 161 |
| PS10 | pIC50 | K1 | 67 |
| CHEMBL730080 | pEC50 | K1 | 989 |
| CHEMBL896244 | pED50 | 3D7 | 230 |
| CHEMBL896245 | pED50 | K1 | 201 |
| CHEMBL1038869 | pEC50 | SB-A6 | 163 |
| CHEMBL1038870 | pEC50 | D10 | 160 |
| CHEMBL730081 | pEC50 | 3D7 | 168 |
| CHEMBL730641 | pEC50 | K1 | 162 |

Table 4.1: 17 subsets of ActivityDB. Where the set consists purely of ChEMBL data, its ChEMBL ID is given as the subset name.

### 4.2.1 Target annotations from original articles

Most of the annotated compounds come from TCAMS set - 833 unique molecules are explicitly named as binders to specific targets, allegedly causing their antimalarial activity. It should be noted that for the majority the target annotations are hypothetical, although of a different degree of confidence.

293 of them have not been associated with a *Plasmodium*-specific protein: while they were found active against the parasite, there is no clear indication towards a target, thus the original paper has used known human targets for these compounds. A human target has only been considered if it was enriched by active compounds, with the enrichment factor being the ratio of target actives among hits, over a background defined as all target actives among all screened compounds with data for that target. Such targets include GPCRs, nuclear hormone receptors, ion channel, *etc*. For all compounds of this category the confidence label is set to "Enriched human target".

Other compounds have been annotated by targets found in orthology studies. Orthology searches for homologous sequences in genomes of different organisms, descended from a common ancestor. In this case, human – *P. falciparum* sequence homology was studied. Most of found targets are kinases, but several specific antimalarial (dihydroorotate dehydrogenase, cytochrome $bc_1$) or antibacterial (dihydrofolate reductase and tRNA synthesis related) targets are also considered. These molecules are labeled by "Orthology"

confidence level.

The most prominent family of targets in the TCAMS set is kinases, and two types of kinases are distinguished: Ser/Thr protein kinase, and $Ca^{2+}$/calmodulin dependent protein kinase. Annotations for aspartic protease, cysteine protease, dihydroorotate dehydrogenase, dihydrofolate reductase and DNA gyrase are kept in the combined data table. All compounds targeting tRNA synthetases are labeled "tRNA synthetase". Cytochrome bc1 inhibitors are annotated as "Electron transport chain" in the original article. Other considered targets include agonists/antagonists of various GPCRs (labeled "GPCR"), nuclear receptors ("Nuclear hormone receptors"), "Ion channel" inhibitors, and some other less prominent enzymes for which the annotations are kept as-is.

ActivityDB contains molecules from St. Jude hospital screening set (chemblid), and for 20 of them the original paper [162] proposes experimentally tested modes of action. They are hemozoin formation inhibition (13 molecules), dihydroorotate dehydrogenase (3), dihydrofolate reductase (2), cytochrome $bc_1$ (2). For all compounds of this set the confidence label is set to "Experimental".

MMV is the most diverse set from the point of view of considered targets. 150 compounds are annotated, but 25 of them are not experimentally validated, but labeled as "structurally related" or "similar" to a certain series, in which case the confidence label is set to "Similarity". As a general rule, the target annotation are left as-is except for kinases and hemozoin formation related molecules.

### 4.2.2 Experimental target annotation from ChEMBL database

All molecules from the MalariaDB set that have a "Hypothetical" mode of action annotation have been looked up in ChEMBL to see if there are experimental validation of considered target. The script was written in Python and considered following rules:

- Target organism for bioactivity record is "Plasmodium";
- Target name for the record is **not** "Plasmodium". Otherwise phenotypic assays are taken into account;
- Activity comment (special field in ChEMBL bioactivity record) says "active".

If the record for a molecule follows these rules, and target type is the same as in the hypothesis, its confidence level is set to "Experimental", and information about assay is stored in the table.

In total, 102 molecules have been found to have experimental evidence of target that way. Most of them target kinases (12 for $Ca^{2+}$/calmodulin dependent protein kinase and 84 for other kinases), 2 were found active against dihydrofolate reductase and 4 against dihydroorotate dehydrogenase.

Next, all non-annotated molecules have been checked for the same experimental validation, but in that case the target was taken as it is, since no hypothesis to check is present. The rules are slightly modified:

- Target organism for bioactivity record is "Plasmodium";
- Target name for the record is **not** "Plasmodium";
- Activity comment (special field in ChEMBL bioactivity record) says "active" *or* measured activity value corresponds to binding or dose-response (EC50, IC50, AC50, $K_i$) and is inferior to 1000 nM.

The modification to the last rule is made since many assays have no unambiguous annotation of active and inactive compounds, but provide instead the numerical value of activity.

154 compounds were annotated that way. Again, most of them are inhibitors of kinases (85 molecules), but other targets are also present: hexose transporter (24), heat shock protein (24), dihydrofolate reductase (9), protein farnesyltransferase (7), dihydroorotate dehydrogenase (2) and 1 example for histidin-rich protein, glucose-6-phosphate dehydrogenase and M1-family aminopeptidase.

The summary of all target annotations is shown in Figure 4.2.

### 4.2.3 Target grouping

We did the target grouping according to the following strategy. Combination of molecules with different targets into one group has been done based on bibliographic research: molecules allegedly hitting target implied into the same pathway are grouped together and labeled as "pathway-specific". Additionally, targets of one family are merged (so, all inhibitors of kinases, whatever the type, are combined). The number of compounds corresponding to each group is shown in Figure 4.3. If groups corresponding to one target were too small, or their classification and merging into a certain mode of action proved to be difficult, they were not labeled, but kept to provide negative ('inactive') examples.

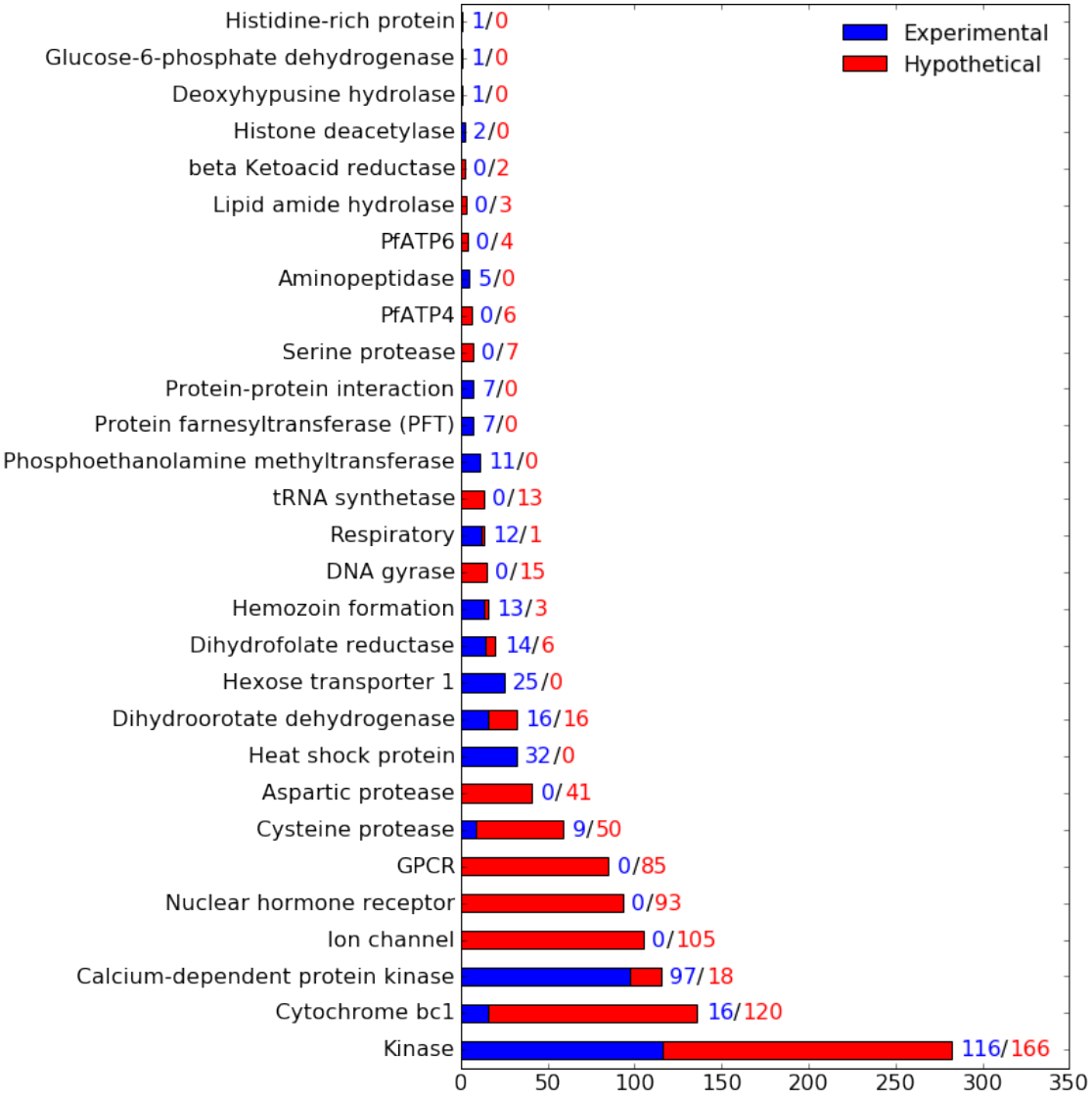



FIGURE 4.2: Total number of molecules for all targets in three datasets of MalariaDB. Number of compounds with experimental (experimentally tested against a given target, in blue) and hypothetical (annotations based on similarity, orthology or human targets, in red) target annotations is indicated for each target.

***Kinases*** is the most prominent group in all datasets. A compound annotated active against any kind of kinase is assigned to this group. These include Ser/Thr protein kinases and $Ca^{2+}$/calmodulin dependent protein kinases (CDPK) from TCAMS and other ChEMBL assays, mitogen-activated (MAPK), cyclin dependent kinases (CDK) and protein kinases G (PKG) from MMV. Interestingly, although Ser/Thr protein kinases are annotated separately in the TCAMS set, other considered classes are also examples of serine-threonine specific kinases.

***GPCR***, ***Ion channel***, and ***Nuclear hormone receptor*** follow their assignment from the original paper [119].

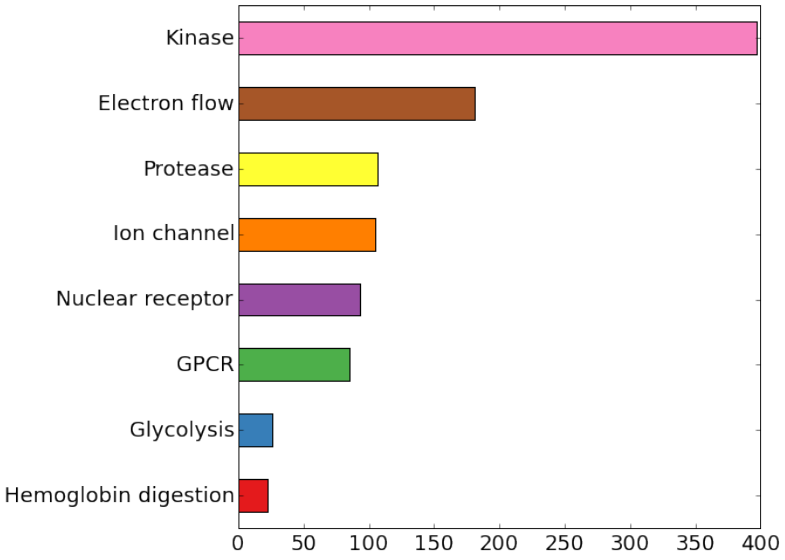



FIGURE 4.3: Number of molecules in all considered target groups.

***Electron flow*** group contains molecules with targets Cytochrome $bc_1$ (cyt $bc_1$), Dihydroorotate dehydrogenase (DHODH) and Respiratory (target annotation from MalariaBox [123]). All those are considered to interfere with the electron flow in the parasite, but at different stages – DHODH enters in earlier stages of redox metabolism and is responsible for pyrimidine metabolic cycle, while cyt $bc_1$ regulates the electron transport chain in mitochondria. "Respiratory target" is a general annotation in MMV for molecules that were found active in high/low oxygen conditions, and is also considered to participate in cell respiration and redox-related processes.

***Glycolysis*** group contains mostly Hexose transporter (HT) inhibitors [163] (numerous in experimental data from ChEMBL) and the inhibitor of glucose-6-phosphate dehydrogenase [164].

***Hemoglobin digestion*** group unites molecules that inhibit the crucial part of the parasite's food cycle. One type of compounds in this group are the inhibitors of hemozoin formation that prevent the crystallization and detoxification of hemoglobin [21]. The others inhibit aminopeptidases, interfering in later stages of digestion [165].

## 4.3 Data standardization and descriptors generation

All compounds used have been standardized following the default protocol installed on the public web server of the Laboratory of Chemoinformatics, and powered by ChemAxon [166, 167] tools. The standardization scheme includes removal of very large entities

(>100 heavy atoms), counter-ion strip-off, split-charge representation of N-oxides, basic aromatization, conversion to the most populated microspecies of the most probable tautomeric form at pH=7.4, etc.

39 diverse ISIDA [126] fragmentation schemes have served to generate molecular descriptors for all compounds. ISIDA/Fragmentor2015 [168] software have been used. The types of fragments include sequences, atom pairs, circular fragments and triplet counts, with information on atoms and/or bonds, colored by atom symbols, pharmacophore features or force field types. These 39 fragmentations were selected for their relatively low number of fragments they generate (less than 3000 each).

# Chapter 5

# Towards a universal map

Cartography (from Greek χαρτης *khartes*, "papyrus, sheet of paper, map"; and γραφειν *graphein*, "write") is the study and practice of making maps. A map is a symbolic depiction emphasizing relationships between elements of some space. The most familiar type of maps for us is the geographical map that depicts the Earth and all spatial information, with the representation of continents and oceans, forests, mountains, plains, and rivers. But, as the definition suggests, a map is not only useful in geography. It shows the relationships of objects of any space, and the chemical space can be the object of cartography. A new term has been proposed by T. Oprea [157] specifically for the purpose of the cartography of chemical space:

> "We therefore suggest the term *chemography*, by analogy with geography, as the art of navigating in chemical space.
>
> The objective of chemography is to provide a consistent mapping device ... that can avoid extrapolations when positioning the properties of a new arbitrary collection of lead-like or drug-like organic molecules."

The key point here is the comparison of chemography to geography. The major particularity of the geographical map of the world is its universal character. Whatever feature one wants to be mapped, the core of the map will not change - it is still the same combination of continents and oceans (Figure 5.1). Even though the world is ever-evolving, new species of plants and animals are found, sea level or temperature change, the contours of the map stay the same and accommodate the changes.

In chemistry, the situation is different. So far, mapping was mainly a problem-specific task: a map is being specifically built to support some specific working hypothesis with

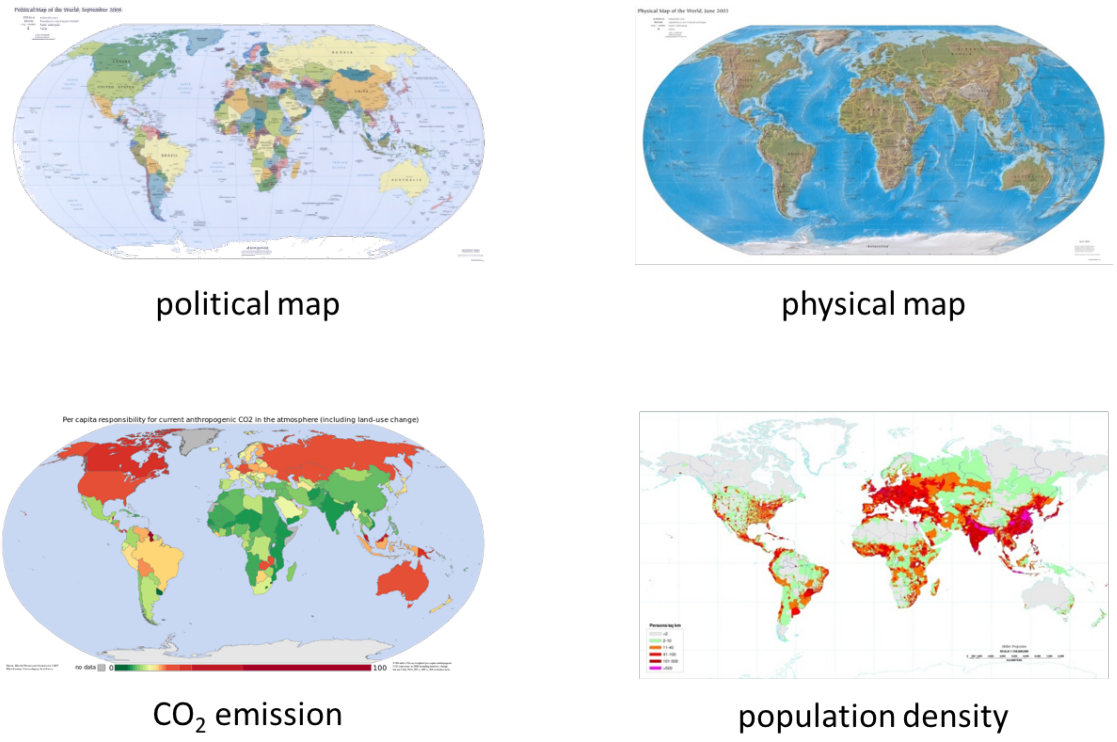



FIGURE 5.1: Examples of the geographical map of the world, colored by different properties (`http://www.nationsonline.org`)

respect to some given compound sets. Maps are considered satisfactory if they are consistent with *a priori* knowledge. For instance, PCA has been used to delineate the CS of drug-like molecules from evidently non-drug-like compounds [157], thus the criterion was the effective clustering of drug-like against non-drug-like CS. Scaffold trees are often used to generalize knowledge about key structural motives present in molecules having particular biological activity, as well as to investigate potential new scaffolds [169].

Often, the mapping of CS is done with drug discovery as an objective: anti-cancer activity [170], oral availability [171], environmental properties [172] or general protein affinity [173]. It serves to identify low-populated zones, or to prioritize molecules mapped in zones populated by compounds of interest. Successful clustering of active compounds associated to different targets is routinely used as a quick, visual validation of a map quality [174]. Other works propose methodologies to interconnect two distinct spaces: ligand CS and the biological space of targets, either by Ligand Efficiency Indices [175] or mapping properties directly as axes [176] in 2D representation. However, these works have focused on a limited number of molecules and molecular properties, and each activity is modeled separately.

On the other hand, the ultimate goal of creation of a map that encompasses all possible (and impossible) chemical structures, along with their molecular, biological, or physicochemical properties, is not yet achieved. So-far realized maps of large virtual compound

sets are often difficult to interpret. It is difficult to rationalize a strategy to design an interpretable and informative map based on unlabeled compounds. Even if modern computer science supports exhaustive enumeration and mapping of billions of virtual low-weight compounds [177, 178], these remain a negligible fraction of the alleged $10^{33}$ drug-like compounds of up to 36 heavy atoms [179]. Since there is no experimental information associated to such virtual compounds, there is no objective way to judge the relevance of the proposed maps. Also, the number of compounds needed to build a useful map (further on denoted as the ***frame set***, as they outline the reference frame spanned by the map) may be less relevant than their diversity. There are reports [180] that the frame set size does not correlate with the SOM quality (defined by their ability to focus on the CS zone relevant for the similarity-based screening). The right combination of frame set size and diversity must be found.

The key question is "What is a good CS map?", and it is still open to discussion. The analogy to geography would suggest the principal criterion of a map claiming to be a universal representation of chemical space: the ability to project novel compounds, including those significantly different from ones used to train the map, in such a way as to ensure that the resulting projection is in compliance with the neighborhood behavior (NB) principle [181, 182], irrespective of the monitored ligand properties. This compliance should be ensured by a quantitative criterion, which would evaluate, in the case of drug-like compounds, the polypharmacological competence of the map.

In addition to the definition of a universality criterion, the issue of creation of such a map remains. It is mentioned above that the definition of chemical space by frame set and descriptors has a tremendous influence on the resulting map. Additionally, sophisticated mapping methods such as GTM have a number of internal parameters that may change a map significantly. This chapter describes the influence of GTM parameters and meta-parameters on the map, and introduces the core methodological finding of the project – a strategy of building an optimal universal map of chemical space.

## 5.1 GTM parameters and meta-parameters

### 5.1.1 Dataset

The influence of method parameters on the resulting map is illustrated on an arbitrary dataset of molecules. This set consists of 100000 randomly chosen molecules from ChEMBL [160] with molecular weight lower than 500 (this was chosen to exclude the abundance of polypeptides as well as to facilitate the following fragmentation; on the other hand, some of molecules had too small molecular weight to produce fragments, so

they were excluded, too). The molecules were standardized, which led to some losses, and the final number of molecules was 99978.

For this dataset, a set of ISIDA descriptors – sequences of atoms and bonds of length 3 to 7 – has been calculated. In total, 19452 descriptors were generated.

### 5.1.2 Effect of descriptor space on the map

The results of dimensionality reduction depend strongly on the initial choice of descriptors. Since different descriptor types convey different information on a compounds and highlight different structural or physico-chemical features, the algorithm would try to fit a function following this specific information. So, in other words, this changes the axes along which the chemical space will be spanned. This is especially critical in the case of activity cliffs, because the cliffs may appear or not depending on our definition of the descriptor space. That influence has been illustrated on the DUD dataset [183]: different descriptor types have resulted in different data points distribution over the map and influenced the clustering score (Figure 5.2).

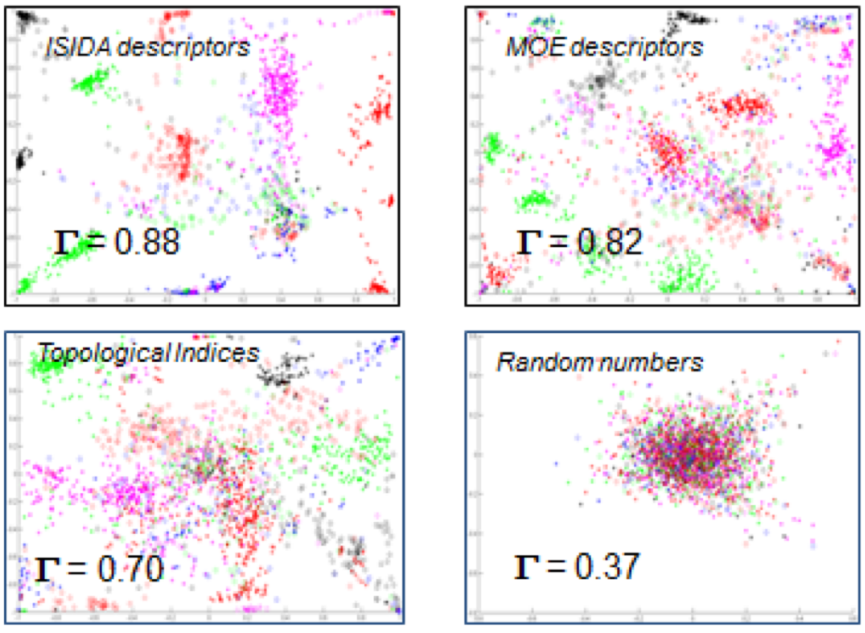



Figure 5.2: Examples of GTM maps built on the DUDS dataset using different descriptors: ISIDA, MOE, topological indices, or random numbers [152].

GTM algorithm is no different in this regard, and we do still face the dimensionality curse problem. Often, the optimization of GTM manifold for a high-dimensional descriptors space generates a map where data points tend to aggregate on manifold nodes,

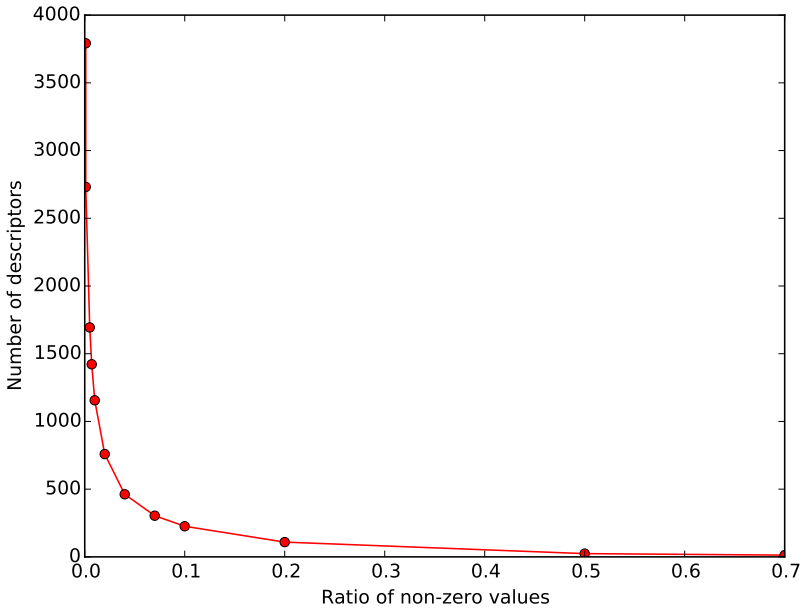



FIGURE 5.3: Number of descriptors kept as opposed to the value of parameter $p$ (minimum ratio of non-zero values). The initial pool of 19452 descriptors reduces to 1156 at $p = 0.01$ and to 20 at the strictest value of $p = 0.7$

while smaller descriptors space dimensions make molecules to be distributed more uniformly on the map. Also, maps with higher number of descriptors require more computational time. Since the dimensionality of the ISIDA descriptor space used in the work is often very high, its reduction would be reasonable for both lower computational time, better look, and, in the ideal case, interpretability of the map.

The obvious way to reduce the number of dimensions is the direct choice of descriptors. To do so, a strategy already implemented in ISIDA/GTM program has been applied: a special parameter $p$ allows to specify the ratio of molecules having non-zero value of a descriptor, and the descriptors which have more null values than needed are discarded. Thus, we lose some of information on specific features, but reduce significantly the descriptor space.

A scale of values for the parameter $p$ from 0.001 to 0.7 has been studied (Figure 5.3). For $p = 0.001$ (descriptors having less than 0.1% of non-null values) almost 4000 descriptors are kept, while for the strictest cleaning (allowing very frequent descriptors with 70% of non-null values) only 20 descriptors are kept. As it can be seen from the figure, even for a rather strict 1% threshold almost 1000 descriptors stay.

Next, several maps were built using reduced descriptors spaces, with otherwise same parameters and incremental algorithm ($25 \times 25$ nodes, $5 \times 5$ RBFs of width 2, regularization

parameter is 1). They were visualized to demonstrate the distribution of molecules on the map (Figure 5.4).

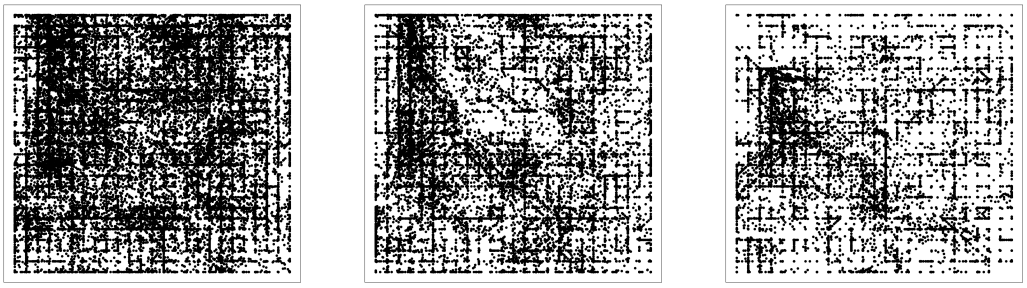


FIGURE 5.4: Maps built on different-sized descriptor spaces ($p$ is 0.1, 0.04, 0.01, from left to right). The more descriptors there are, the more molecules tend to aggregate onto the manifold nodes.

The visualization of the map is not always sufficient to demonstrate the distribution of molecules on the map, since black points represent only the mean position of a molecule having a probability distribution over the map. To complement, entropy [149] of molecules (a value that shows the distribution of the responsibility: 0 means it is concentrated on one node, 1 – distributed uniformly all over the map) was calculated by following formula:

$$H = \left( -\sum_k R_k \log R_k \right) / \log K. \tag{5.1}$$

Here, $R_k$ is the responsibility of a molecule on $k$-th node, $K$ is the number of nodes. The sum is normalized in range of 0 to 1 (therefore divided by $\log K$). Histograms for each case (Figure 5.5) show that as the value of $p$ decreases (thus, the size of the descriptor space increases) the entropies of molecules tend to converge to zero, *i.e.* the distribution of responsibilities becomes more unimodal.

### 5.1.3 Effect of manifold flexibility

It is also interesting to study the effect of manifold flexibility on maps, since it can affect significantly the look and the modeling ability of a map. Two parameters of ISIDA/GTM are directly related to the manifold flexibility – the number of RBF centers and the regularization coefficient. Both of them were studied with respect to the distribution of molecules on the map.

1. **Regularization coefficient** $\alpha$. Maps with different value of regularization coefficient (from 100 to 100000 with factor of 10) were built and visualized for the same

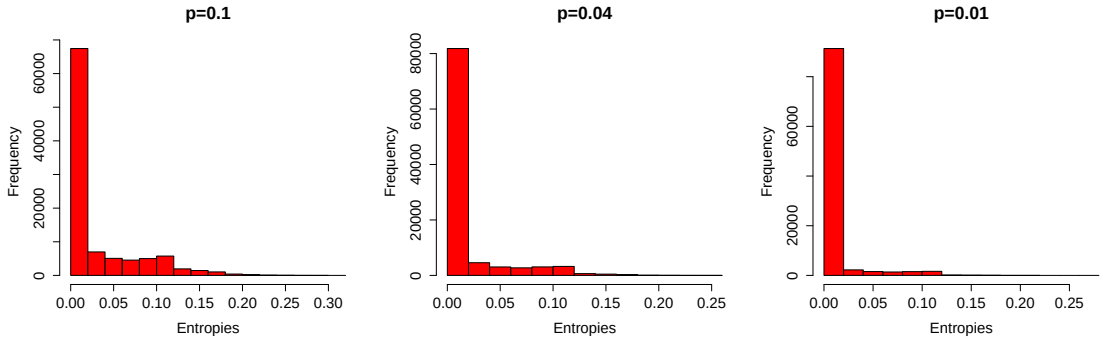
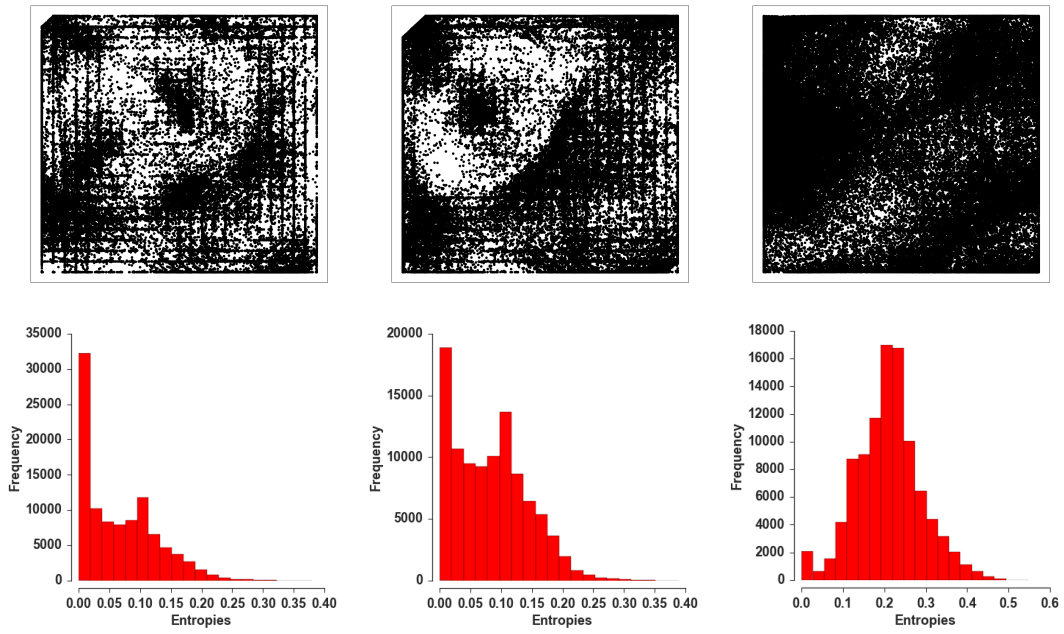



FIGURE 5.5: Histograms of entropies of molecules for maps built on desciptor spaces of different sizes (different values of $p$). As the number of descriptors decreases, entropies tend to grow, *i.e.* responsibilities become less unimodal.

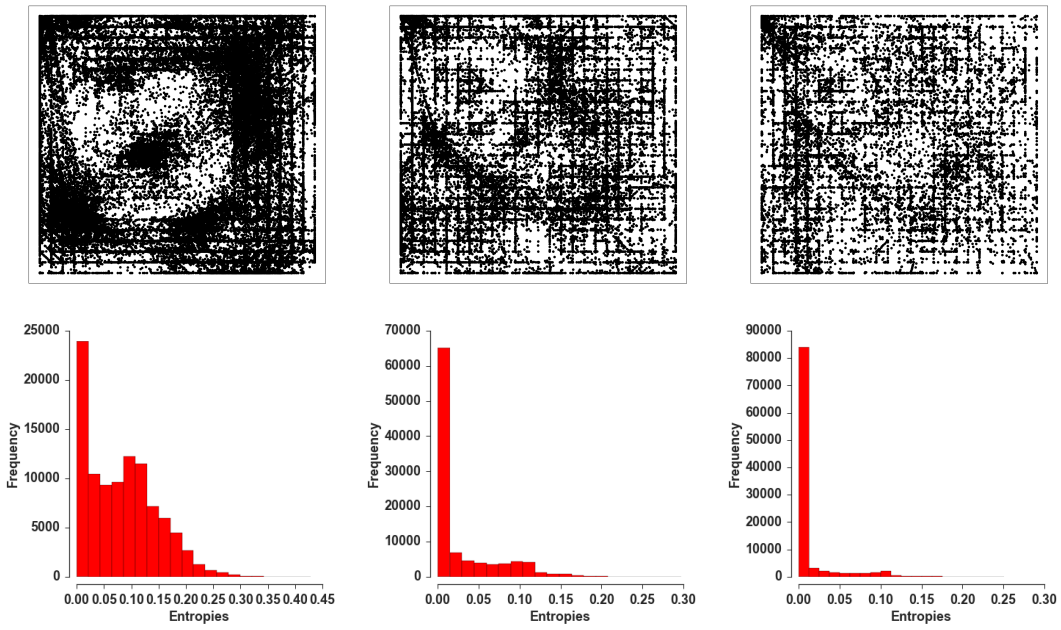
subset as before, with cleaned descriptor space (only descriptors with more than 4% non-zero values were considered, $p = 0.04$). All other GTM parameters are kept constant: the map resolution is $25 \times 25$ nodes, RBF grid is $5 \times 5$, RBF width is set to 2. As the regularization coefficient increases, the loglikelihood $\mathcal{L}$ of the model decreases very insignificantly. The entropy of molecules was also calculated. The visualizations and histograms (number of molecules having certain entropy) are shown below (Figure 5.6).

As it can be seen, the more rigid the manifold is, the more uniformly the molecules




(A) $\alpha = 100$ (B) $\alpha = 1000$ (C) $\alpha = 100000$

FIGURE 5.6: Maps and entropy distributions for different values of regularization parameter $\alpha$. The higher the $\alpha$, the more rigid is the manifold, and molecules tend to have more uniform responsibility distribution over the map.




(A) $4 \times 4$ (B) $8 \times 8$ (C) $14 \times 14$

FIGURE 5.7: Maps and entropy distributions for different sizes of RBF grids. The same tendency is observed: the more rigid is the manifold (the smaller the grid of RBFs), the more molecules have smooth responsibility distribution over the map.

cover the latent space, but the responsibility of molecules becomes more delocalized over the map.

2. **Number of RBF centers**. To the same effect, the influence of the number of RBF centers was studied. Maps on the same dataset with the same descriptors and regularization coefficient 1, the same map resolution as before, with different number of RBF centers (grids from $4 \times 4$ to $14 \times 14$) were made. Their visualizations and histograms (on the same principle) are shown on Figure 5.7.

In both cases, the maps follow the same trend : the more flexible is the map (higher number of RBF centers or lower $\alpha$), the more the molecules tend to converge to nodes of the manifold, since the manifold adapts itself very closely to data. On the other hand, for rigid manifold (low number of RBFs or high $\alpha$) the molecules, as well as their responsibility, are more distributed over the map. At the most extreme case of a rigid manifold the method would convert to a kind of a non-linear PCA: molecules would be projected on a 2D plane through RBF functions. On the other hand, extremely flexible and big manifold would transform the resulting model into SOM.

## 5.2 Map optimization

As described above, the map depends on meta-parameters, such as frame set and descriptor type, and internal method parameters like map resolution, manifold flexibility through regularization parameter and RBF characteristics. Although common sense would suggest limits for each option, some of them take continuous value, and complete enumeration of possible maps is beyond feasibility.

Previously, map optimization in ISIDA/GTM tool was only implemented *via* batch procedure [184]. This process consists in systematically building maps with varying parameters and evaluating it in QSAR classification or regression model. While this method is convenient for a smaller pool of considered parameter setups, it has several drawbacks. The first one is, as stated above, the impossibility of complete enumeration of setups. Second, no clear relation between model quality and parameter values has been yet established. Third, this procedure has only been limited to one property and small datasets.

The advantage that the GTM proposes is the peculiar way of building a GTM-based QSAR model. The modeling is divided in two steps: building a manifold in the first stage and actual quantitative model training in the second. Since manifold can be trained on one set of data (frame set) and evaluated by any other, GTM can actually be proposed as a tool for building a universal map. In this case, frame set size and composition, descriptors, and all parameters can be optimized to produce a map that is capable of accommodating and predicting an array of molecules and properties that may not be related.

Batch optimization is not applicable in that case due to sheer number of possible map candidates. Here we propose a genetic algorithm (GA) [185] based map optimization procedure. GA is an optimization technique widely used in engineering [186], chemoinformatics [187, 188], artificial intelligence [189] and other domains. It is inspired by the process of the natural evolution, which has reflected in the terminology of the method.

The parameter tuning strategy is based on another one that has been reported earlier for SVM method [190]. It is schematically represented on Figure 5.8, and is explained in detail further.

### 5.2.1 Input

The parameters of a candidate GTM are encoded by a ***chromosome***. It is a vector of specifications needed to build this map. Part of this chromosome contains the GTM

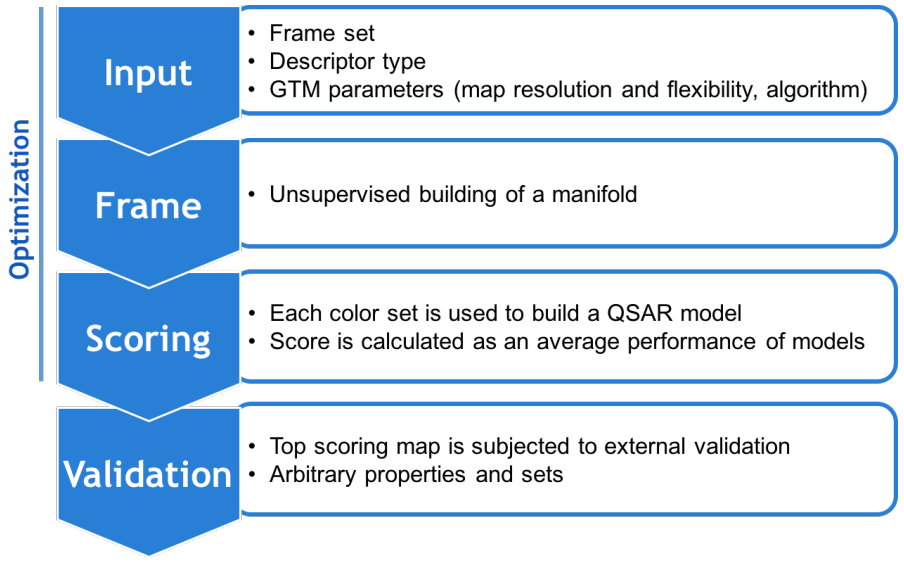



FIGURE 5.8: Schematic representation of GA-based optimization workflow. Each step is explained in details in the text.

setup: number of nodes, number of RBF functions defining the manifold and their width, the regularization coefficient [147]. The other parts of the chromosome are the type of molecular descriptors and the frame set to use. These define the studied chemical space and, in the broader sense, the applicability domain of the map. One last part encoded by a chromosome is the prediction method used in the scoring step. Possible options are explained in section 3.4.1.

### 5.2.2 Manifold construction

The algorithm reads, from the current chromosome, the encoded key parameters that are relevant for the manifold building: choice of the frame set, of the type of descriptors, size, number of RBFs, etc. The manifold construction is an unsupervised process: experimental property values associated to compounds play no role in the fitting of the GTM manifold. The role of the frame set compounds required at building stage is to span relevant CS zones, and thus define a reference frame for the manifold, not to provide any property-related information.

### 5.2.3 Scoring

Map quality assessment consists in building GTM-driven regression or classification models for each of the considered ***color sets***. Color sets are small activity-annotated datasets, encoded in the same descriptor space as the given map requires. For each

color set $s$, a cross-validated QSAR model is constructed, and then a set-specific cross-validated determination coefficient $R_s^2$ or balanced accuracy $BA_s$ is returned. The cross-validation procedure is repeated $N_{trials}$ times (three by default). The herein defined universality criterion or map fitness score is the mean of all set-specific values of $R_s^2$ or $BA_s$ . Additionally, maps producing near-equal set specific success scores are preferred over maps returning some very high and others very low values. Therefore, the fitness score of the GA was taken as following for regression mode:

$$Score =< R_s^2 > -0.5 \times \sigma(R_s^2), \tag{5.2}$$

or, for classification tasks:

$$Score =< BA_s > -0.5 \times \sigma(BA_s). \tag{5.3}$$

Here, $\sigma$ denotes standard deviation of all set-specific scores.

Other types of scoring functions have also been used by others. For example, if compounds of the set do not bear property annotations, the map can be guided to adapt in a way that accommodates the most chemical entities. In this case, the objective function would be the loglikelihood $\mathcal{L}$ or the minimum variance of $\mathcal{L}$ between frame and external sets. Alternatively, clustering characteristics, entropy or other parameters can be used.

### 5.2.4 Genetic algorithms

Genetic (or evolutionary) algorithm is a stochastic optimization algorithm whose heuristics resemble those of the natural selection. At each iteration of the algorithm a population of possible solutions (in this case, possible combinations of GTM parameters and meta-parameters, or, so to say, the map recipe) is generated. These represent "genotypes" of future maps. Manifolds built according to these recipes are the "phenotypes". Each individual manifold is assigned a "fitness" score based on how successful they were in modeling the color sets.

In the evolutionary algorithm formalism, the "fitness" of a given manifold determines its chances to be selected and have offspring (by cross-overs and mutations) to transfer the knowledge of the quality of the setup to produce "universally" competent maps. This characterizes the method's heuristic for the selection of the best map.

### 5.2.5 External validation

In order to challenge and evntually confirm or discard the claim of 'ùniversal" applicability of these maps, they should be exposed to an external validation. As stated above, the properties associated to the external sets may, and should, differ from those associated to the color sets. However, the applicability domain of the map should be taken into account, a map trained to model properties of small organic drug-like molecules will not be applicable for the properties of polymers in material science study.

## 5.3 Universal map of drug-like space

As a proof-of-concept study, a challenge of construction of universal map of drug-like space has been taken up.

### 5.3.1 Frame sets

Five candidate datasets were selected for the frame.

- *Set1*: the diverse set [180] of marketed drugs, biological reference compounds, ligands from PubChem database, as well as randomly picked ZINC [191] compounds;
- *Set2*: a subset of the ChEMBL dataset provided by Prof. Bajorath, where only one-third of ligands of each target are taken. If the resulting number of compounds selected for a target was less than 50, then the target was discarded. Ligands appearing in several target set are represented only once, leading to a frame set of 9877 entries;
- *Set 3*: a subset of above-mentioned, where half of ligands for half of targets are taken. In the same manner as for Set2, only target which contributed at least 50 compounds were selected. In total, there are 7214 unique entries;
- combinations of *Set2* and *Set3* with *Set1*, *e.g.* two fused set labeled *Set1+2* and *Set1+3*, respectively. Duplicates were kept only once.

### 5.3.2 Coloring sets

The dataset used at map selection stage contains 165 ChEMBL compound series, all larger than 50 compounds curated and provided by Prof. J. Bajorath, University of Bonn.

Set members are all the compounds with reported $pK_i$ values with respect to the associated targets (receptors, enzymes, etc.), without any particular choice of targets. In total, 49989 data entries are present.

Out of the 165 individual target-related ChEMBL sets, 21 appear to be non-modelable by SVM with any of the 39 considered descriptor types (fitness score < 0.5). Therefore, only the remaining 144 sets served as coloring sets.

### 5.3.3 External validation sets

In order to validate the universal character of map candidates, several challenges were set up including both the chemical information within the coloring sets and external molecules and properties (with difficult to model *in vivo* properties).

An overview of all external validation sets is displayed in Table 5.1. Challenges 1 and 2 concern the same subsets of ChEMBL as used for building and selection of maps, while other challenges use external datasets, i.e. considering both molecules and biological properties that are completely unrelated to any of the compounds and binding affinities used for map building and selection process. Challenges 3 and 4 are the global target-specific activity classification challenges which concerns all possible targets other than the ones in the selection sets. The data collection and annotation was done through an automated mining and curation procedure.

### 5.3.4 Results and discussion

Eventually, five top maps corresponding to five descriptor spaces were found. The parameters and descriptor types of the maps, numbered in decreasing fitness order, are presented in Table 5.2. They were selected as candidates for the universality challenge.

The observations of the maps' parameterization are the following. First, the preferred frame sets are the *Set2* and *Set3*, which are the samples of ChEMBL. Adding independent *Set1* compounds in any combination did not improve the model, since maps based on Sets 1, 4, or 5 did not demonstrate top performance. Second, pharmacophore- and force-field-colored atom pair and atom triplet counts were much preferred over more sophisticated sequence and augmented atom types (that are only present in Map 1, which was not as performant in external validation challenges). This is not entirely surprising since the map is supposed to be relevant for drug-like space, thus, taking into account pharmacophore features or force-field type (which is a more detailed atom classification scheme) is more consistent with medicinal chemistry interests.

| Set name | Number of properties | Description |
|---|---|---|
| Challenge 1 | 144 | Discriminating active and inactive for each one of the 144 targets of coloring sets |
| Challenge 2 | 5 | Classification of ligands by target family: kinases, nuclear receptors, peptidases, monoamine GPCRs, other GPCRs (coloring sets) |
| Challenge 3 | 410 | Active/inactive classification task, covering human single protein targets of ChEMBL. Active compounds have an associated dose-dependent activity value ($K_i$, IC50, EC50) lower than an activity threshold |
| Challenge 4 | 468 | The Challenge 3 dataset enriched with compounds with reported "potency" values |
| Challenge 5 | 20 | Classify as inhibitor/non-inhibitor and substrate/non-substrate of transmembrane transporters [192] |
| Challenge 6 | 17 | Classify as active/inactive antimalarial candidates (ActivityDB) |
| Challenge 7 | 7 | Classify as active/inactive antivirals for 7 viral types [155] |

Table 5.1: External validation sets used to verify the universal map candidates in classification tasks.

| Map | Descriptors | Frame set | Resolution |
|---|---|---|---|
| 1 | IIRAB-PH-1-2: pharmacophore-colored augmented atoms: sequences of atoms and bonds of fixed length, covering first and second coordination sphere | Set3 | $40 \times 40$ |
| 2 | IAB-FF-P-2-6: force-field-type-colored counts of atom pairs, 1 to 5 bonds apart, including information on bonds nearest to terminal atoms | Set2 | $32 \times 32$ |
| 3 | IA-FF-P-2-6: same as above, but without bond information | Set3 | $39 \times 39$ |
| 4 | IAB-PH-P-2-14: pharmacophore-colored counts of atom pairs 1 to 5 bonds apart, including information on bonds nearest to terminal atoms | Set2 | $32 \times 32$ |
| 5 | III-PH-3-4: pharmacophore triplets, with edges of topological distances 3 and 4 | Set3 | $40 \times 40$ |

Table 5.2: Descriptors types, frame sets, and map resolutions for top five maps. For descriptors, ISIDA nomenclature [168] is given. Map resolution reflects the size of the square latent space grid.

All five maps perform well in external validation (Figure 5.9). Map 2 has demonstrated overall higher performance in all challenges, but not by much. All maps are competent to capture the rather fuzzy structural signatures that define ligand families (monoamine and other GPCR, kinase, peptidase, and nuclear receptors, for Challenge 2). Excellent results were also obtained with respect to Challenge 3. More than 90% of properties were predicted with BA > 0.6. Challenge 4 was significantly more difficult for all maps,

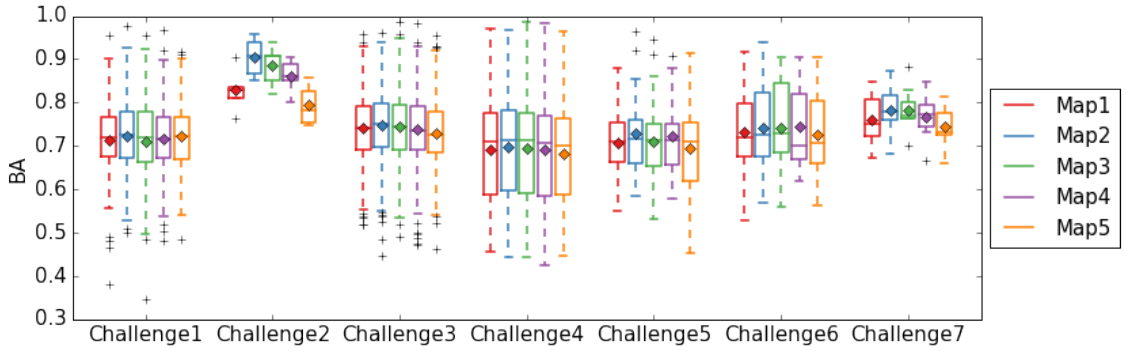



FIGURE 5.9: Boxplots of performances of top 5 maps in 7 external validation challenges. Performance is measured in balanced accuracy (BA). Mean BA for the challenge is indicated by a diamond.

although not surprisingly, considering large amount of noisy data from HTS studies with reported "potency" values.

The selected maps also handle the classification of *in vivo* properties in very satisfactory way. Map 4 is especially competent with respect to separation of active and inactive compounds in antimalarial activity datasets. Challenge 7 concerns the classification of antivirals active against one family of viruses against all other families, and is solved especially well by maps 2 and 3 (minimal BA is 0.68 and 0.70, respectively).

The maps are also useful for the analysis of the chemical space corresponding to a certain target or target family and their comparison. For example, the comparison of chemical space coverage for Cox-2 inhibitors and Cytochrome 2C9 inhibitors leads to reasonable results: coloring the map by activity against 2C9 results in a denser and more homogeneous landscape, since many compounds are screened against it. The landscape of Cox-2 inhibitors is more pointed, since the medicinal chemistry studies focus often on certain known series of anti-inflammatory agents.

As an additional example, the distribution of compounds on Cox-2 ligands landscape has been studied more closely (Figure 5.10). For the analysis, the Cox-2-associated compounds were assigned to key medicinal chemistry series: coxibs, fenamic acids (-fenacs), -profenes, indometacin derivatives, and more recently discovered series of benzopyrans [193] and tetrahydrobenzofurans [194]. The actives of some families (indomethacin derivatives and foremost benzopyrans) are focused onto a small number of nodes. On the other hand, diverse coxibs and tetrahydrobenzofuranes are non-homogeneously spread over a significant map area. Nevertheless, within the broadly spread coxib-like series, locally homogeneous subfamilies can be clearly localized on the map, and their localization is driven more by substituents than the classical for medicinal chemistry understanding scaffold.

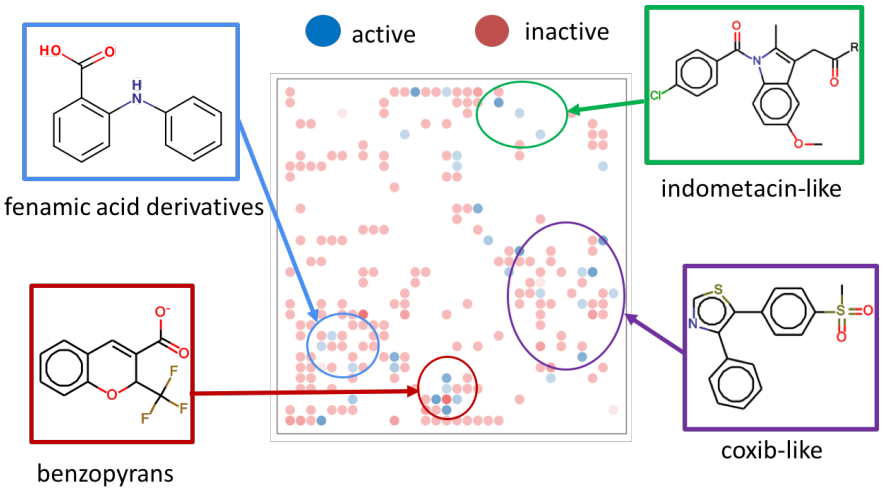



FIGURE 5.10: ACM representation of Cox-2 (CHEMBL230) ligands on the Map2, cross-validated balanced accuracy = 0.7. Red nodes are mostly populated by inactives, blue ones - by inactives. Color intensity represents the data density of the node. In the zoomed-in portions – common substructures for ligands

The chemical space coverage has also been studied by the comparison of responsibility distribution of targets of different families on the map. An overlap of chemical space coverage of related targets can be detected on the map, which is reasonable from the medicinal chemistry point of view. For example, adenosine receptor inhibitors are quite separated from one another, logically, since they belong often to different chemical series. Most adenosine receptors inhibitors [195, 196] are, in fact, derivatives of adenosine, theophylline, or xanthine, and are selective binders. On the other hand, many kinase inhibitors are common between proteins due to their high promiscuity [197, 198]. An example of chemical space density analysis between a target and target family is shown on Figure 5.11. Here, for G-coupled protein receptors (GPCRs), four hierarchical levels were highlighted: dopamine D1, dopamine family, monoamine (rhodopsin-like) GPCRs, and all GPCRs. It can be seen that the dopamine family is already quite representative of the rather homogeneous monoamine GPCR binders, whilst the monoamine GPCR subspace is a limited subdomain of the entire cluster of GPCR ligand zones. Since the manifold for the universal map is the same, the landscape even for unrelated targets may be compared.

## 5.4 Conclusion

This part of the thesis addressed the question whether a Universal, compound set-independent Generative Topographic Map can be generated, with the universality claim quantitatively justified.

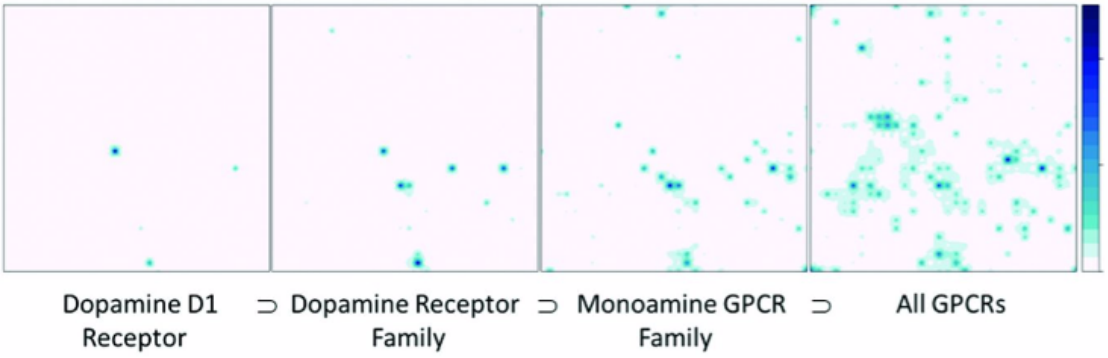



FIGURE 5.11: Density plots of associated ligands represent target, or target-family specific signatures on a GTM. Density is represented by cumulative responsibility of compounds of a given class. Supported by a "Universal" map, the signatures of the different targets can be quantitatively compared. [147]

The proposed strategy of the universal map creation is novel because it does not match either of two dominating approaches used in chemoinformatics for polypharmacological prediction: classical multi-target QSAR and chemogenomics. Like in classical QSAR, the approach generates one individual model per target. It does not require, like in chemogenomics, target-specific descriptors. However, unlike classical QSAR, which requires tedious fitting of each individual model, the key advantage of our approach is that all tunable parameters and meta-parameters (descriptor choice) are already determined, and were shown to represent a good choice for a plethora of classification models of biological properties (target binding and *in vivo* activities) which were completely unrelated to properties used for fitting. In other words, models are still target-specific (one needs to project, for each property to model, a dedicated structure–property data set onto the map), but their parameters are not.

The peculiar property of the GTM algorithm – the manifold is built in an unsupervised manner, and then colored by property, by mapping of a dataset with associated experimental labels – is at the core of proving the Universal character of evolved maps. The maps are perfectly suited to solve classification problems concerning chemical structures never used to fit the map: on the overall, more than 80% of the more than 600 distinct and varied classification problems, chosen such as to cover a maximum of exploitable SAR data, were successfully solved.

In addition, the maps provide with an intuitive representations of the data. They were shown to provide a consistent analysis of the drug-like space. This strategy and these GTM models will, in our opinion, be quite helpful to provide a better global overview of modern trends and challenges in medicinal chemistry.

J Comput Aided Mol Des (2015) 29:1087–1108
DOI 10.1007/s10822-015-9882-z

# Mappability of drug-like space: towards a polypharmacologically competent map of drug-relevant compounds

**Pavel Sidorov[1,2] · Helena Gaspar[1] · Gilles Marcou[1] · Alexandre Varnek[1,2] · Dragos Horvath[1]**

Received: 5 October 2015 / Accepted: 6 November 2015 / Published online: 12 November 2015


**Abstract** Intuitive, visual rendering—mapping—of high-dimensional chemical spaces (CS), is an important topic in chemoinformatics. Such maps were so far dedicated to specific compound collections—either limited series of known activities, or large, even exhaustive enumerations of molecules, but without associated property data. Typically, they were challenged to answer some classification problem with respect to those same molecules, admired for their aesthetical virtues and then forgotten—because they were set-specific constructs. This work wishes to address the question whether a general, compound set-independent map can be generated, and the claim of "universality" quantitatively justified, with respect to all the structure–activity information available so far—or, more realistically, an exploitable but significant fraction thereof. The "universal" CS map is expected to project molecules from the initial CS into a lower-dimensional space that is neighborhood behavior-compliant with respect to a large panel of ligand properties. Such map should be able to discriminate actives from inactives, or even support quantitative neighborhood-based, parameter-free property prediction (regression) models, for a wide panel of targets and target families. It should be polypharmacologically competent, without requiring any target-specific parameter fitting. This work describes an evolutionary growth procedure of such maps, based on generative topographic mapping, followed by the validation of their polypharmacological competence. Validation was achieved with respect to a maximum of exploitable structure–activity information, covering all of *Homo sapiens* proteins of the ChEMBL database, antiparasitic and antiviral data, etc. Five evolved maps satisfactorily solved hundreds of activity-based ligand classification challenges for targets, and even in vivo properties independent from training data. They also stood chemogenomics-related challenges, as cumulated responsibility vectors obtained by mapping of target-specific ligand collections were shown to represent validated target descriptors, complying with currently accepted target classification in biology. Therefore, they represent, in our opinion, a robust and well documented answer to the key question "What is a good CS map?"

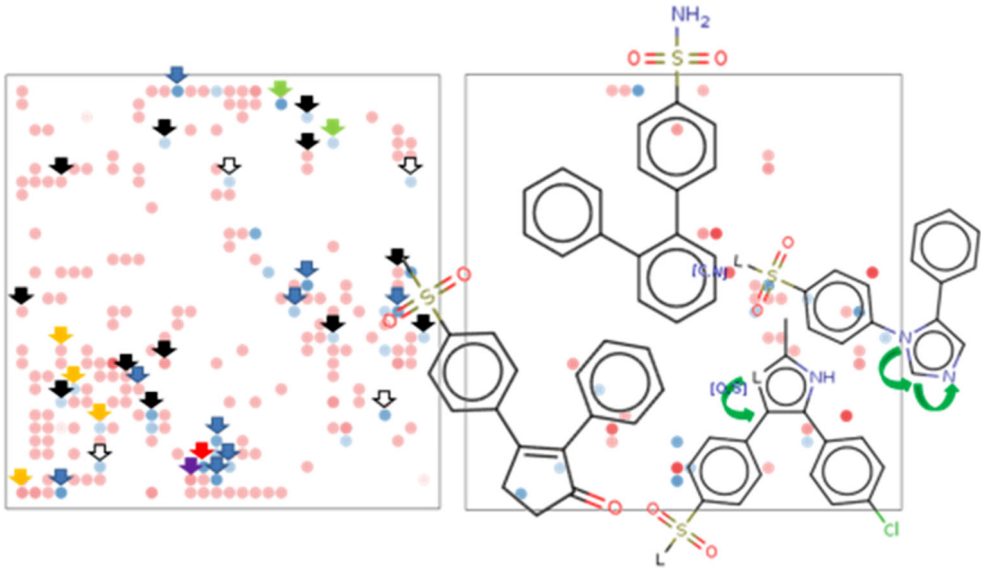
**Electronic supplementary material** The online version of this article (doi:10.1007/s10822-015-9882-z) contains supplementary material, which is available to authorized users.

✉ Dragos Horvath
dhorvath@unistra.fr

[1] Laboratoire de Chémoinformatique, UMR 7140, CNRS-Univ. Strasbourg, 1 rue Blaise Pascal, 67000 Strasbourg, France

[2] Laboratory of Chemoinformatics, Butlerov Institute of Chemistry, Kazan Federal University, Kazan, Russia

**Graphical Abstract**




**Keywords** Structure–property relationships · Polypharmacology · Chemical space mapping · Generative topographic maps

**Abbreviations**

| | |
|---|---|
| (Q)SPR/SAR | (Quantitative) structure–property/structure–activity relationships |
| CS | Chemical space |
| GTM | Generative topographic map |
| HTS | High throughput screening |

## Introduction

Quintessentially, a map represents a simplified, lower-dimensional model of reality, capturing, with minimal distortion, the objective neighborhood relationships between mapped items. Mapping of the planetary sphere surface onto a rectangular frame can be achieved by various projection methods, but, irrespectively of the one used, the same map can then be colored/interpreted by political, geophysical or economic criteria. Interpreting the map in terms of any novel property—say, average Internet use of populations—is perfectly feasible on maps drawn well before Internet age. Also, the emergence of a new city does not require map reconstruction, but mere projection of the new item within the already given frame.

In chemistry, it is of high practical interest to emphasize chemical space (CS) zones associated to favorable molecular properties [1]. Therefore, CS mapping is an area of active research in chemoinformatics. Medicinal chemists visualize [2] the "islands" of binders to a receptor, embedded in the "ocean" of inactive structures, and try to conclude on specific features of each island, in terms of structural traits of their ligand "inhabitants". CS mapping implies—unlike in geography—extremely aggressive dimensionality reduction, resulting in unavoidable distortion of the initial CS. The challenge of CS mapping is to find the dimensionality reduction procedure specifically preserving the problem-relevant chemical information.

Mapping approaches in chemoinformatics are extremely popular, and there is a wealth of dimensionality reduction algorithms—linear principal component analysis (PCA), and non-linear approaches—Self-Organizing Maps [3, 4] (SOM), Multidimensional Scaling [5] (MDS), Stochastic Embedding [6], 2D scaling with rubber bands [7], Generative Topographic Maps [8, 9] (GTM).

So far, mapping was mainly a problem-specific task: a map is being specifically built to support some specific working hypothesis with respect to some given compound sets. Maps are considered satisfactory if they are consistent with a priori knowledge. For instance, PCA has been used to delineate the CS of drug-like molecules from evidently non-drug-like compounds [10], thus the criterion was the effective clustering of drug-like against non-drug-like CS. Scaffold trees are often used to generalize knowledge about key structural motives present in molecules having particular biological activity, as well as to investigate potential new scaffolds [11].

Often, the mapping of CS is done with drug discovery as an objective: anti-cancer activity [12], oral availability [13], environmental properties [14] or general protein affinity

[15]. It serves to identify zones of low density, or to prioritize molecules mapped in zones populated by known actives. Successful clustering of actives associated to different targets is routinely used as a quick, visual validation of a map quality [16]. Other works propose methodologies to interconnect ligand CS with the biological space of targets, either by Ligand Efficiency Indices [17] or mapping properties directly as axes [18] in 2D representation. However, they focus on a limited number of molecules and molecular properties, and each activity is modelled separately.

Alternatively, so-far realized maps of large compound virtual compound sets are difficult to interpret. It is difficult to rationalize a strategy to design an interpretable and informative map based on unlabeled compounds. Even if modern computer science supports exhaustive enumeration and mapping of billions of virtual low-weight compounds [19, 20], these remain a negligible fraction of the alleged $10^{33}$ compounds of up to 36 heavy atoms [21]. Since there is no experimental information associated to such virtual compounds, there is no objective way to judge the relevance of the proposed maps. Also, the number of compounds needed to build a useful map (further on denoted as the 'frame' set, as they outline the reference frame spanned by the map) may be less relevant than their diversity. There are reports [3] that the frame set size does not correlate with the SOM quality (defined by their ability to focus on the CS zone relevant for the similarity-based screening). The right combination of frame set size and diversity must be found.

The key question 'What is a good CS map?' is still open to discussion. The analogy to geography would suggest the major criterion to be fulfilled by a map claiming to be a universal representation of chemical space: the ability to project *novel* compounds, (including those significantly different from ones used to train the map), in such a way as to ensure that the resulting projection is in *compliance with the neighborhood behavior* (NB) principle [22, 23], irrespective of the monitored ligand properties.

Or, this is hardly the case so far, as can be concluded from the above-mentioned state-of-the-art. It is not obvious that a map properly delimitating islands of active compounds with respect to a target $T$ will also be able to serve for monitoring the distribution of actives and inactives of another target $T'$. One reason for such failure may be map training, limited to a frame set of molecules spanning some finite CS volume relevant for ligands of $T$, but not of $T'$. Another, equally important, is that $T$-compatible molecular descriptors may not capture $T'$-relevant chemical information. However, if NB-compliance with respect to $T'$ is observed, then this map is one step closer to the ideal of "Universal CS map", compared to all alternative mapping schemes that fail $T'$. While this ideal may be out of reach, the goal of this paper is to actively seek for mapping strategies maximizing NB-compliance, over a maximum of unrelated biological targets and properties.

This is achieved by introducing a quantitative universality criterion as a measure of NB-compliance in a vast, polypharmacological context, as will be detailed later on. As a side note, it is important to emphasize that a strict definition of NB compliance is of paramount importance to uphold the claim of maximal map quality: obtaining distinct "islands" of steroids and benzodiazepines is a trivial task achieved by any basic structure pattern recognition tool, whilst discriminating between active versus inactives steroids, and, respectively, active versus inactive benzodiazepines is the type of hard problem used in present benchmarking.

Figure 1 below depicts the CS mapping paradigm, with both its key issues and working hypotheses that are central to the herein advocated strategy.

The practically infinite number of possible CS maps stems from a combination of huge pools of choices for each of the critical elements of the mapping process (*item 1* in Fig. 1).

- First, as already discussed, a map must be built on hand of a *set of frame compounds* that span the relevant CS zones. Typically, in problem-specific map building, this frame set is identical to the set of compounds to analyze. In the present quest for a universal rather than property-bound mapping strategy, the choice of frame compounds (out of several predefined options) was integrated as an explicit degree of freedom.
- Second, the *choice of descriptors* is the core of the CS definition. The quest for a universal mapping strategy cannot rely on the typical "educated guess" of the descriptor type best suited to solve a specific problem, but will consider descriptor choice as another explicit degree of freedom (out of a wealth of diverse, chemical information-rich ISIDA [24–26] property-labeled fragment counting schemes).
- Third, mapping success will depend on the chosen *algorithm* and its peculiar *setup*. Currently, the algorithm choice was restricted to generative topographic mapping (GTM) [8, 9, 27–29]—in particular, the incremental GTM algorithm [28]. Nevertheless, the specific GTM parameters need to be chosen: a GTM renders a compound by fuzzy mapping on a square grid (size?) of nodes, after having optimized a flexible manifold described by Gaussian functions (how many? what width?).

Even with above-mentioned restrictions, complete enumeration of possible maps as conceptually stated in the Fig. 1 is beyond feasibility. The exploration of this huge phase space of mapping options will be undertaken by an evolutionary algorithm [30]. This presumes the introduction

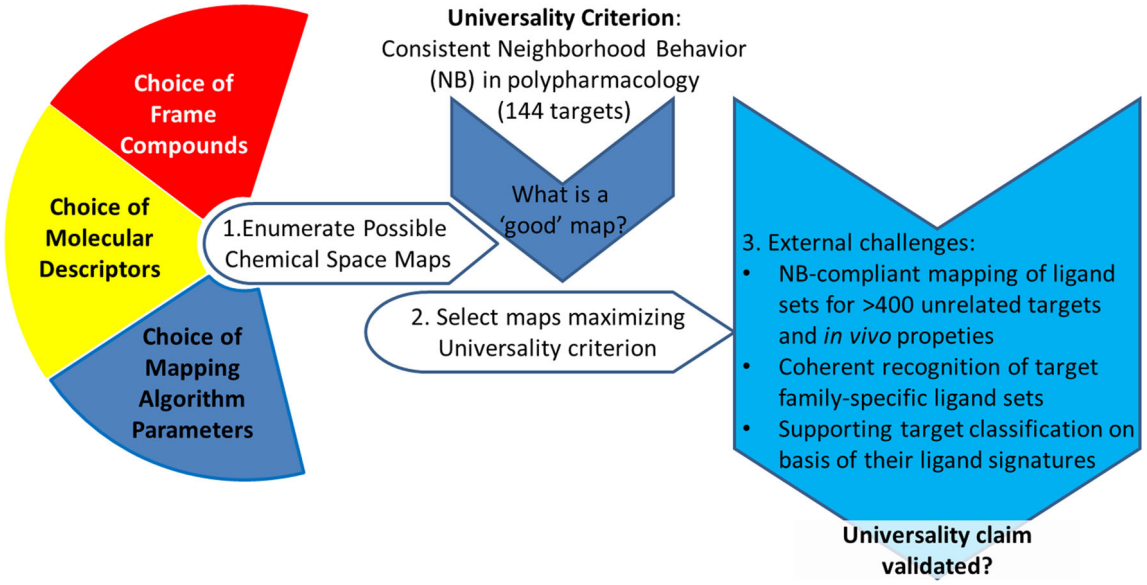



**Fig. 1** Generic paradigm of CS mapping, completed with the specific working hypotheses used in the present quest for a universal map

of a quantitative universality criterion (*fitness score*, *item 2* in Fig. 1) supporting the Darwinian selection of mapping schemes of maximal generality. Fortunately, GTMs support [8, 27] neighborhood-based property prediction, for both continuous (regression) and categorical (classification) properties. The more predictive the model supported by a given map, the better its NB-compliance, thus the better the map—with respect to that property. However, this work does not focus on any specific property, and thus the universality criterion relies on an entire panel of—unrelated and diverse—biological properties. It is defined as an aggressively cross-validated NB compliance measure in a vast polypharmacological context regrouping 144 targets.

Eventually (*item 3* in Fig. 1), the most evolved map candidates were subjected to extensive challenges meant to verify in how far the claim of universal applicability holds, beyond the polypharmacological context used for selection. External testing involved a maximum of exploitable structure–activity information, covering all of *Homo sapiens* proteins of the ChEMBL database, antiparasitic and antiviral data, etc. The three distinct categories of challenges below were designed to be as exhaustive as technically possible, within available computational resources and experimental information:

- In an attempt to ensure that external validation addresses a significant part of published drug-relevant structures of reported activity, the ability to discriminate between active versus inactive ligands of more than 400 biological targets and in vivo activities was assessed by mapping on the top 5 selected universal map candidates. Both external ligands and targets were arbitrarily dissimilar to the selection context.
- Alternatively to active/inactive classifications per targets, maps were also challenged to separate ligands by the target families to which they preferentially bind.
- Additionally, the tool was shown to be able to coherently describe the relatedness of various biological targets based on the ligands associated to them. The idea to directly measure or predict the functional relatedness of receptors by monitoring which pairs of targets coherently display similar levels of affinity throughout a series of common or similar ligands [31–34] is important, because bioinformatics-driven alternatives based on sequence comparison are not accurate enough (diverging sequences may nevertheless hide similar functionality) whereas site geometry-based considerations require structure elucidation and tedious 3D-model driven predictions. Unlike cited approaches, in which a same set of ligands must be tested on both targets in order to assess their mutual functional similarity [31, 34], or else demands tedious pairwise ligand similarity computations [33, 34], the advantage of this method is that for any target with known ligands, it is possible to extract from the map a vector representation of limited dimensionality, accounting for a large simplification of the target–target comparison task.

Results showed that, indeed, maps built following the herein outlined strategy, satisfactorily solved hundreds of activity-based ligand classification challenges for targets, and even in vivo properties independent from training data. They also stood chemogenomics-related challenges, as cumulated responsibility vectors obtained by mapping of target-specific ligand collections are valid target

descriptors, complying with currently accepted target classification in biology. Therefore, in our opinion, they are the closest to the ideal definition of a universal map of the CS.

## Methods

### Generative topographic mapping

Generative topographic mapping (GTM) is a method [29] of non-linear mapping that has been successfully used in various domains of data analysis. The following is a brief reminder of GTM methodology, mentioning the recurring *keywords* in the domain.

In GTM, each point in the low-dimensional (usually 2D) *latent space* (*LS*) is mapped onto the *manifold* embedded in the initial CS. The manifold is defined by a mapping function $y(x; W)$ assessed with the help of $M$ radial basis functions (RBFs) of width $w$ regularly distributed in LS. The latent space is covered by a squared *grid* of $K$ *nodes* ($K$ being a perfect square), each of which corresponding to a normal probability distribution (NPD) centered on the manifold. The NPD is used to compute the *responsibilities* $R_{kn}$, representing the degree of association between the CS point of compound $n$ and the node $k$. Alternatively, $R_{kn}$ is the fuzzy-logics truth value of the statement "Compound $n$ resides in node $k$". Therefore, $\sum_{k=1}^{K} R_{kn} = 1$.

Note that responsibilities $R_{kn}$ of compounds $n$, members of a specific library/subset (binders to a given target or family of targets, for example), can be summed up to obtain *cumulated responsibility vectors*, characterizing the whole compound collection (and, by extrapolation, the associated target, if the compound collection is target-specific). In the present context, *mean cumulated responsibility* vectors will be used exclusively (therefore henceforth omitting the "mean" label), where the sum of compound responsibilities is divided by the number of contributing compounds. A (mean) cumulated responsibility vector is technically an object identical to any compound responsibility vector, i.e. denotes the fuzzy mapping of the compound collection over the GTM nodes, with its elements summing up to 1.0. It may be used in neighborhood studies (similarity scoring) like any other single-compound vector. Cumulated responsibility vectors can be understood as descriptors of the *CS coverage* of a compound library. This feature will be used to compare targets represented by the cumulated responsibility vectors of their ligands.

The introduction of real-value responsibility vectors, i.e. fuzzy probabilities of simultaneous association of a compound to—potentially—several nodes at the time, is the key difference between GTMs and classical Kohonen SOMs. The latter may formally be regarded as based on a binary responsibility vector: $R_{kn} = 1$ for a single node $k$ where the compound $n$ is said to "reside", and $R_{jn} = 0$ for any $j \neq k$. All residents of a node have a same binary "responsibility" vector and are thus, de facto, indistinguishable entities on a SOM. Therefore, the maximal number of distinct clusters/structural families that may be resolved on a SOM cannot exceed the total number of grid nodes $K$. On a GTM, a compound may be shared between different nodes, with different probabilities, and any of such distinct "sharing schemes" may represent the signature of a specific cluster/structural family. Formally, the number of distinct states that may be encoded by a map is given by the phase space volume of its responsibility vector. For a binary vector, it equals $K$. Counting the—a priori much more numerous—distinct states in a continuous responsibility phase space empirically depends on the size of the envisaged unit phase space cell (occupied by sets of near-redundant analogs). Also, not the entire phase space is chemically relevant: compounds will be typically distributed with significant probabilities over a few neighboring nodes, not equally smeared, with low probability, over all nodes at a time. Therefore, it is not possible to determine a priori how many nodes would be needed to properly host all clusters/families of drug-relevant compounds.

GTM regression models [27] are relying on the neighborhood principle. First, training molecules are located on the map, and GTM nodes are 'colored' by the responsibility-weighted mean values of input experimental property values of training compounds. Next, the left-out (test) subset is also projected, and test molecules get their predicted properties assigned from the reference values of the neighboring nodes. This general process can be carried out under different premises—basically different formulas to calculate neighborhood weights, and therefore may, for a same manifold and a same training/test set, return different predictions.

GTM classification models [8] follow the same principle, the difference being that nodes are colored by dominating class, and test molecules are then classified by kNN or Bayesian approach. In this work, Activity Class Maps (ACM), rendering the distribution of a two-class (active/inactive) compounds collection on the map, will be of paramount importance in result analysis and visualization. The ACM is an image of the square grid of GTM, where color intensity of a node reflects the normalized summed-up responsibilities of that node for all the mapped molecules. This is a measure of local compound density. Color choice reflects the dominant class of the node—a node is said *active* (herein colored in blue) if the sum of

responsibilities of that node for all the active compounds exceeds the one over the inactives.

### Map generation, selection and validation: key concepts

This paragraph provides a general overview of the herein used evolutionary strategy of searching a best candidate for a putative universal map of drug-relevant chemical space, and introduces specific names for the many protagonists—various data sets, algorithms, etc.—involved. Detailed specification of all these will follow.

The parameters of a candidate GTM are encoded by a *chromosome*. It is a vector of specifications needed to build this map. The flowchart in Fig. 2 shows how map quality—*fitness score*—is estimated, starting from the encoding chromosome.

Part of this chromosome contains the GTM setup: number of nodes, number of RBF functions defining the manifold and their width, the regularization coefficient [8, 9, 27, 28]. Another part of the chromosome is the *type of molecular descriptors* to use. The map can be built on various molecular descriptor types, to encompass diverse ways to represent chemical information. The optimal chemical description strategy emerges from the Darwinian evolution process to which map chromosomes are submitted.

GTM construction is an unsupervised process: experimental property values associated to compounds play no role in the fitting of the GTM manifold. The role of the 'frame' compounds required at building stage is to span relevant CS zones, and thus define a reference frame for the manifold, not to provide any property-related information. There is no straightforward way to define a frame compound sample guaranteeing the production of good quality maps. Therefore, several alternative frame sets were provided as an explicit degree of freedom encoded chromosome. It is left to Darwinian evolution to select the ones leading to enhanced quality maps. The constitution of frame sets will be detailed in the Data section below.

The last locus in the chromosome encodes the modality to use the map as a regression model. It describes several alternative implementation GTM regression protocols [27]. This locus has no incidence on the generated manifold, but different usage in regression modeling will lead to different assessment of the fitness score of the GA.

Map assessment consists in building GTM-driven regression models for each of the considered structure–activity sets. These will further on be denoted as *Selection Sets* in order to distinguish them from the *Challenge Sets*. For each selection set $s$, the given manifold operating in the chromosome-prone *PredMethod* mode produces a list of cross-validation predictions $P_{pred}$. Predictions are confronted to the experimental property values $P_{exp}$ (here, enzyme and receptor inhibition $pK_i$ values). A set-specific cross-validated determination coefficient $Q^2_s$ is returned. The cross-validation procedure is repeated $N_{trials} = 3$ times, thus for each set there are three $Q^2_s$ values. The herein defined *universality criterion* or map *fitness score* is

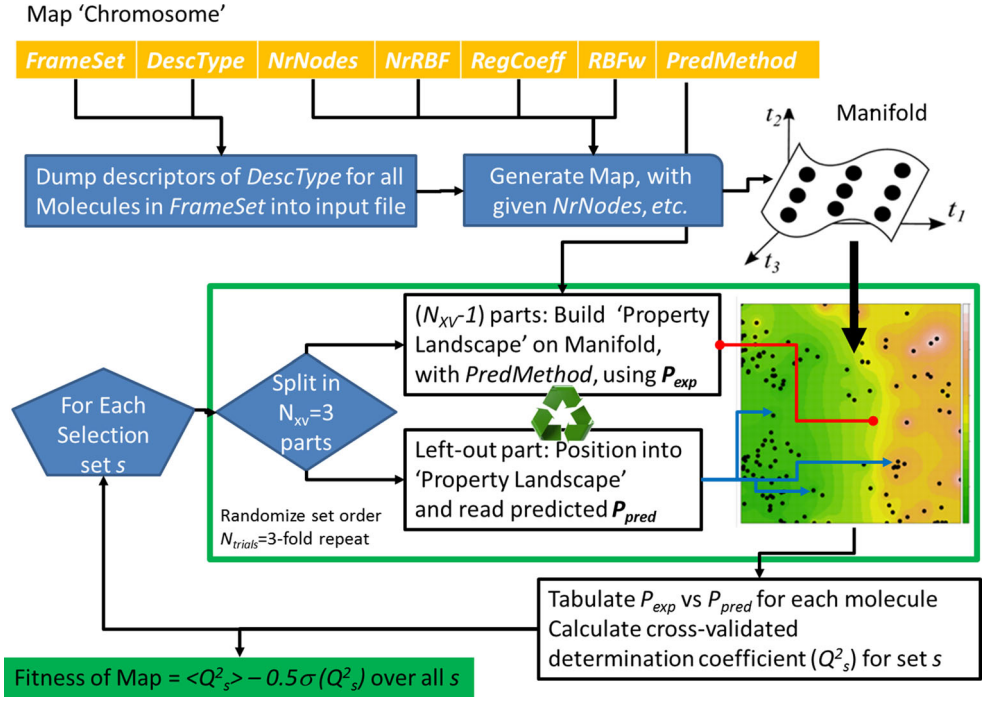



**Fig. 2** Given a setup (frame set, molecular descriptor type and method parameters,) encoded in a *chromosome*, a GTM manifold (a map candidate) is generated. For each selection set $s$, a threefold cross-validated GTM regression model is built, and the cross-validation determination coefficient $Q^2_s$ is tabulated. The map *fitness* (universality criterion) is then calculated as the mean value of $Q^2_s$ penalized by its variance over all $s$

the mean of all set-specific coefficients $Q_s^2$. If the number of sets is $S$, then the number of averaged $Q_s^2$ values is $3S$. Maps producing near-equal set specific success scores are to be preferred over maps returning some very high and others very low $Q_s^2$ values. Therefore, the fitness score of the GA was taken as the mean $\langle Q_s^2 \rangle$ penalized by (0.5 ×) its standard deviation $\sigma(Q_s^2)$.

In the evolutionary formalism, the manifold corresponds to the phenotype "incarnating" its chromosome, and its fitness determines its chances to be selected and have offspring (by cross-overs and mutations) in an asynchronous, distributed GA [30].

As mentioned, *Challenge Sets* are distinct compound collections of properties differing from those covered by *Selection Sets*, and used in the last phase of this work. They are meant to externally validate the maps fitted on the selection sets. However, there is a precise reason for which these were called "challenge", rather than "validation" sets. In QSPR, external validation means properly predicting properties of external compounds. The nature of the herein envisaged "challenges" is fundamentally different—it is to verify whether the independently constructed map is able to properly accommodate the new compound sets within the absolute frame it stands for. An external challenge set is properly accommodated if the map succeeds to serve as support for a valid QSPR model thereof. Hence, the challenge set—unlike a "validation set" in the classical acceptation of this term—will provide the property data needed to both *train* and cross-*validate* this QSPR model. Iteratively, two thirds of the challenge set will serve to "color" the map by the associated (categorical) property ad build the associated ACM, i.e. to *train* the QSPR model. The remaining tier plays eventually the role of the classical external QSPR *validation* set: its molecules are blended into the ACM and classified as actives and inactives. The default "kNN" method was used in all classification challenges [8]. After completion of a cycle of three iterations, each tier of the challenge set has eventually played its role as validation set, each challenge set molecule owns a predicted class value. This cycle is repeated three times, using a randomized splitting into three tiers. The final predicted class value is thus the one returned in at least two of the three repeats. Predicted and actual compound classes are compared, and the balanced accuracy *BA* (defined as the mean of specificity and sensitivity) is calculated. Since predictions are issued from a cross-validation scheme, we herein propose—by analogy to the cross-validated determination coefficient $Q_s^2$—to denote the above-defined cross-validated BA score by $QBA_s$. The % of challenge sets $s$ achieving a cross-validated balanced accuracy $QBA_s > x$ can be plotted on $y$ against this variable threshold $x$, producing curves that always start at (0, 100 %) and should decrease as slowly as possible with $x$ (ideally remain at 100 % until $x$ approaches its maximum 1.0). A larger ratio of well modeled individual properties achieving high $QBA_s$ (higher area-under-curve) means stronger support in favor of the "universality" hypothesis.

## Molecular descriptors

Compound standardization followed the default protocol [35] installed on our public web server, and powered by ChemAxon [36, 37] tools. It includes removal of very large entities (>100 heavy atoms), counter-ion strip-off, split-charge representation of *N*-oxides, basic aromatization, conversion to the most populated microspecies of the most probable tautomeric form, etc.

A pool of 39 diverse ISIDA fragmentation schemes [24, 25, 38, 39] served as initial choices for the best suited descriptor types for a universal map. They include sequences, atom pairs, circular fragments and triplet counts, colored by atom symbols, pharmacophore features or force field types. More details are provided in the file *DescRules.cmd* in Supporting Information. These 39 fragmentation schemes were selected for their relatively low number of fragments they generate. The detailed fragment composition of each ISIDA fragmentation scheme is available in the *RefMols* subfolder of Supporting Information.

## Data

The main, but not exclusive data source used in this work is the ChEMBL database [40]. A first subset—mainly used at map selection stage, see item 2 of Fig. 1—contains 165 ChEMBL compound series, all larger than 50 compounds curated and provided by Prof. J. Bajorath, University of Bonn. Set members are all the compounds with reported $pK_i$ values with respect to the associated targets (receptors, enzymes, etc.), without any particular choice of targets. In total, 49,989 data entries are present. The file *targets.name_tid* in Supporting Information describes this collection, and the sets of ligand in SMILES format are available upon request.

External challenge datasets (refer to item 3 of Fig. 1) were retrieved, unless otherwise noted below, from the entire ChEMBL (v.20). About 1.28 million of unique chemical entities successfully passed standardization, corresponding to 1.34 million distinct compounds ChEMBL IDs. This collection was sampled to extract challenge sets.

### *Frame sets*

Five different combinations of molecules were used as frame set candidates:

(a) *Set1*: the diverse set [3] of marketed drugs, biological reference compounds, ligands from PubChem database, as well as randomly picked ZINC [41] compounds;
(b) *Set2*: a subset of the ChEMBL dataset provided by Prof. Bajorath, where only *one-third* of ligands of *each* target are taken. If the resulting number of compounds selected for a target was less than 50, then the target was discarded. Ligands appearing in several target set are represented only once, leading to a frame set of 9877 entries;
(c) *Set3*: a subset of above-mentioned, where *half* of ligands for *half* of targets are taken. In the same manner as for *Set2*, only target which contributed at least 50 compounds were selected. In total, there are 7214 unique entries;
(d) combinations of *Set2* and *Set3* with *Set1*, e.g. two fused set labeled *Set1* + *2* and *Set1* + *3*, respectively. Duplicates and/or degenerate descriptor vectors were kept only once.

### *Selection sets*

Out of the 165 individual target-related ChEMBL sets mentioned above, 21 appear to be non-modelable, in the sense that any attempt to generate SVM models with either of the 39 considered descriptor types failed to discover models of fitness score >0.5. Therefore, only the remaining 144 sets served as *selection sets* for the polypharmacologically most competent maps.

### *Challenge sets*

Several challenges were set up in order to confront the map with (1) a different, complementary analysis of the chemical information within the above-mentioned target-specific subsets and (2) external molecules and properties—including difficult to model in vivo properties.

An overview of all challenges is displayed in Table 1 above. Challenges 1 and 2 concern the same subsets of ChEMBL as used for building and selection of maps, while other challenges use external datasets, i.e. considering both molecules and biological properties that are completely unrelated to any of the compounds and binding affinities used for map building and selection process.

*Challenge 1* The "internal" target-specific classification challenge—i.e. discriminating between actives and inactives within each of the 144 before-mentioned ligand sets associated to various targets. To this purpose, for each compound series tested on a different target, a $pK_i$ cutoff was defined as the integer for which the fraction of affinities better than this cutoff is of roughly 25 %.

*Challenge 2* Target family-driven classification challenge, meant to prove that resulting manifolds have the ability to distinguish between broad classes of ligands associated to customarily defined families of related targets. In particular, the maps were challenged to distinguish typical monoamine GPCR ligands, ligands of other GPCRs, nuclear receptor ligands, kinase inhibitors and peptidase inhibitors. To this purpose, the 144 targets associated to selection sets were first regrouped according to their family. The directory *family-based* in Supporting Information contains the lists of targets merged together within every of above-mentioned family.

*Challenges 3 and 4* The global target-specific activity classification challenge. The success of these challenges is considered as an indicator of the universality of a map. It concerns all possible targets *other* than the ones in the selection sets. They required a complete automated mining and curation of ChEMBL *Homo sapiens*-target related structure–activity information, as detailed in the "Appendix" below.

*Challenge 5* The intestine-blood transporter challenge based on the data set from literature [42]. It features lists of inhibitors/non-inhibitors and substrates/non-substrates of each of the eleven listed transmembrane transporter systems. The property of inhibiting a transporter is distinct from being a substrate thereof, and each is covered by specific challenge sets. For some transporters, only inhibition or only substrate classes are reported, leading to a total profile of 20 distinct transporter-related activities that are covered by this challenge.

*Challenge 6* Malaria challenge: screening results of compound series against *Plasmodium* cultures, using different experimental protocols. It was extracted from the MalariaDB subset of ChEMBL, and completed with in-house results concerning original molecules synthesized and tested in Dr. Davioud-Charvet's laboratory, University of Strasbourg (see article [47] and references therein). For each antiparasitic screening protocol, reported potency values ($pEC_{50}$ or similar) were converted to binary active/inactive classes, as for Challenge 1. Results stemming from different protocols featuring only minimal experimental setup differences were simply merged, in order to generate more robust sets of actives/inactives for each of the different approaches to encode antimalarial activity. When in doubt, data sets associated to different experimental protocols were considered as distinct series, i.e. distinct classification challenges. The resulting 17 sets, each associated to experimental protocols briefly outlined in Supporting

**Table 1** Selection and Challenge sets used to evolve and validate the universal map candidates

| Set name | No. of properties | Type | Description |
|---|---|---|---|
| Selection sets | 144 | Regression | Quantitative prediction of $pK_i$ values for each of the 144 targets shown to be modelable, see 3.4.2 |
| Challenge 1 | 144 | Classification | Discriminating actives from inactives for each one of the 144 targets of selection sets |
| Challenge 2 | 5 | Classification | Classify ligands as inhibitor or not of: *kinases, nuclear receptors, peptidases, monoamine GPCRs, other GPCRs*—Based on the selection sets |
| Challenge 3 | 410 | Classification | Structure–activity sets (active/inactive), covering human single protein targets of ChEMBL. Active compounds have an associated dose-dependent activity value ($K_i$, $IC_{50}$, $EC_{50}$) lower than an activity threshold |
| Challenge 4 | 468 (548)[a] | Classification | The Challenge 3 dataset enriched with compounds with reported "potency" values |
| Challenge 5 | 20 | Classification | Classify as inhibitor/non-inhibitor and substrate/non-substrate of transmembrane transporters: *ASBT, BRCP, MCT1, MDR1, MRP1, MRP2, MRP3, MRP4, OATP2B1, OCT1, PEPT1* [42] |
| Challenge 6 | 17 | Classification | Classify as active/inactive antimalarial candidates, for each of 17 anti-Plasmodium testing protocols—ChEMBL MalariaDB data, enriched with novel putative antimalarials synthesized by the group of Dr. Davioud-Charvet, Strasbourg [43–46] |
| Challenge 7 | 7 | Classification | Classify as active/inactive antivirals, for each of 7 viral types: *enterovirus, hepacivirus, influenza A, lentivirus, orthohepadno-virus, pestivirus, simplex* (ChEMBL) |

[a] Number of actual challenge sets associated to the 468 targets: for some targets with very large ligand sets, the latter were split into several independent challenges

Information file *MalariaProtocols.pdf* were used in the present challenge.

*Challenge 7* Antiviral challenge: it is based on ChEMBL compound subsets reported to display one of the seven antiviral activities reported in Table 1. For each virus family, ChEMBL was mined for actives against representative viruses of that family. The role of inactives is played by the active compounds associated to the six other families, minus promiscuous molecules also assigned to the current family's active class.

## Ligand-related target signatures and chemical space coverage of targets

As mentioned in Introduction, the use of ligand data to indirectly characterize targets has already been proven useful [31–34]—therefore, it is important to show that the herein presented formalism is able to cope with this important target similarity principle: "Targets associated to similar ligand libraries are functionally related". A strong feature of GTMs is the straightforward ability to quantitatively compare the overlap of CS covered by two compound collections, encoded by their cumulated responsibility vectors [28]. This degree of overlap of associated ligand space zones can be simply expressed by a covariance score—here, the Tanimoto index *Tc* [48]—of associated cumulative responsibility vectors. This avoids tedious comparison of each ligand of one set to all the members of the other. Are *related* targets associated to compound sets covering *overlapping* CS zones, as described by cumulated responsibility vectors? Note that the whole sets of tested compounds here used—actives and inactives alike—to characterize a target.

Biology-driven target classification of targets was adopted from the ChEMBL SARFARI projects, covering the two most widely covered target super-families, GPCRs and kinases. Targets previously used in this study—both from selection sets and Challenge 3—were matched by their ChEMBL ID codes and protein accession codes against annotated target lists of both SARFARI subsets. Resulting files *gpcr_meta.txt* and *kinase_meta.txt* in Supporting Information assign individual targets to the specific target subfamilies (or biological families, *in memento* of the biology-driven considerations) as defined in the SARFARI files. The detailed classification scheme corresponding to ChEMBL level 4 (listing the main therapeutically relevant GPCR subfamilies) was used, but only target families of at least 4 members were kept for further analysis. Level 2 and 3 classification schemes of GPCRs—encoding rather generic subfamilies, such as "Monoamine GPCRs", "Short Peptide GPCRs", etc.—were considered too coarse for the present analysis, and ignored. By contrast, with kinases only the coarser levels 2 and 3 allowed regrouping targets into subfamilies of minimum four targets, while level four is too specific.

Practically, this challenge can be formulated as follows: biology proposes a classification of targets into families, based on their functional relatedness. Is the GTM-driven

representation of targets as cumulated responsibility vectors of their associated ligands also supporting the hypothesis that mentioned subfamilies 'cluster' together in terms of CS overlap scores?

Above, "clustering together" should not be understood as an algorithm-dependent outcome of some actual clustering procedure, but in terms of increased intra-family *cohesion* over inter-family *separation*. The cohesion of a family is the mean target distance between family members. The separation of a family is the average distance between members of this family to all other members of distinct families. The distance measure $D(T,t)$ between two targets $T$ and $t$ is the opposite of the Tanimoto index [48]: $D(T,t) = 1 - Tc(T,t)$ based on cumulated responsibility vectors of $t$ and $T$ ligands, respectivey. Formally, if the mean of intra-family distances $\langle D(T,t)\rangle$, $T,t \in F$ (Cohesion) is significantly shifted, according to Student's $t$ test [49], towards lower values with respect to the mean $\langle D(T,t)\rangle$, $T \in F$, $t \in F'$ (Separation) between any target in $F$ and any other non-family member $F'$, then the computational method can be said to recognize the internal cohesion of $F$, hence complying to SARFARI classification. Otherwise, if intra- and inter-family distance distributions are not statistically different (or if—never observed—intra-family distances are longer) then the CS coverage analysis fails to recognize that members of $F$ are functionally related. The Student $p$ value may thus serve as a fuzzy-logic truth level indicator for each of the statements "CS overlap analysis is in agreement with the SARFARI classification of members of $F$ into a same functional family". The lower the $p$ value the better the agreement between SARFARI families and CS overlap analysis.

## Results and discussions

### ChEMBL data curation

Only 410 of 2474 ChEMBL targets entered the Challenge 3. It involves 107,510 distinct chemical entities and covers a total of 248,455 experimental ligand-target associations. Out of these 59,162 are positive (active) examples, the others being experimentally validated inactives.

The 410 targets for which enough data was found covers roughly 1/6 of the total number of *Homo sapiens* single proteins from ChEMBL. Eighteen targets were found to harbor both strict dose–response activity types and less well defined *potency* entries: these enter both challenges 3 and 4, but with different structure–activity sets. There are 40 targets featuring only *potency* scores and they are also those associated to the highest quantity of screening results.

The Challenge 4 data extraction strategy covers thus $410 + 18 + 40 = 468$ distinct targets, roughly 1.47 million of ligand-target association data—out of which $\sim$150,000 represent active examples—and involved 449,859 distinct chemical species, roughly 1/3 of the entire database. The largest structure–activity class sets were split in several sub-problems thus leading to 548 distinct structure–activity sets.

### Characteristics of obtained maps

The asynchronous evolutionary process, on 20 nodes (with $6 \times 86\_64$ cores/node) of the HPC cluster of the University of Strasbourg, over a period of roughly 1 month, i.e. $\sim$10 CPU years was stopped after 5000 distinct GTM setup protocols were visited and assessed. Out of these, map building and assessment was prematurely aborted in roughly 1/3 of the cases, either because of convergence failures at manifold fitting stage or because coloring by property lead to inaccurate fitted values, so that further cross-validation would have been a waste of time. The other cases represent fully assessed mapping hypotheses, for which Fig. 3 reports the fraction of cases achieving given lowest, mean and highest $Q^2$ values, over the 144 selection sets at play. It can be seen that, even with the least meaningful parameter choices, at least one of the data sets marginally passes cross-validation, at $Q^2 > 0.35$. This is not astonishing, since the problem space was delimited to include only ISIDA fragment descriptors based on reasonable fragmentation schemes, and map parameter ranges were reasonably estimated. Bad mapping schemes were visited, but they were mainly bad consensus maps all while maintain some marginal predictive power with respect to few targets. Complete mapping failures, due to absurdly low grid size or due to the choice of irrelevant descriptors were not observed. Reversely, even in the best setup schemes, at least one target failing cross-validation, at $Q^2 < 0.05$ can be found. Evolution, mainly driven by the mean $Q^2$ (the corrective term related to its standard deviation being left out of this analysis, for simplicity) can be seen to reach rather quickly the main basin of near-optimal solutions at $\langle Q^2\rangle = 0.25\ldots0.35$.

Eventually, five descriptor spaces were found to be at the basis of the top fitness GTMs. The five maps, numbered in decreasing fitness order, are, each, the fittest with respect to its associated descriptor space. They were selected as candidates for the universality challenge—see Table 2 for a detailed list of parameters, and Supporting Information file *FiveBestMaps.xlsx* for details.

#### *Impact of the frame set choice*

The evolutionary history of the maps showed that the preferred frame sets are the Set2 and Set3, which are

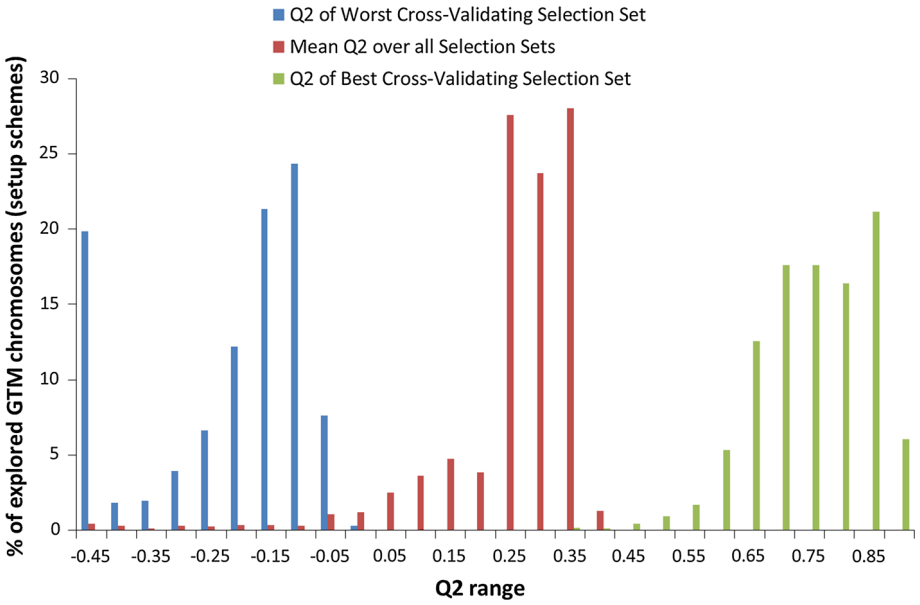



**Fig. 3** Distribution of cross-validated determination coefficients of lowest, mean and highest $Q^2$ over the 144 selection sets at play, for all the valid GTM setup hypotheses encountered during the Darwinian evolution process

**Table 2** Top five maps, each relying on a distinct descriptor space

| Map | Descriptors [39] | FrameSet | Size | NrRBF | RBFw | RegCoeff |
|---|---|---|---|---|---|---|
| 1 | IIRAB-PH-1-2: pharmacophore-colored atom centered fragments based on sequences of atoms and bonds of fixed length, covering first and second coordination sphere | Set3 | 40 | 16 | 1.0 | 8.91 |
| 2 | IAB-FF-P-2-6: CVFF Force-field-type-colored counts of atom pairs found at 1–5 bonds apart, including interposed bond information | Set2 | 32 | 19 | 0.9 | 0.2298 |
| 3 | IA-FF-P-2-6: as above, but without bond information | Set3 | 39 | 17 | 1.1 | 0.0028 |
| 4 | IAB-PH-P-2-14: pharmacophore-colored counts of atom pairs found at 1–5 bonds apart, including information on bonds nearest to terminal atoms | Set2 | 32 | 17 | 0.6 | 0.5754 |
| 5 | III-PH-3-4: pharmacophore triplets, with edges of topological distances 3 and 4 | Set3 | 40 | 15 | 0.2 | 0.3388 |

Size refers to the number of nodes defining the edge length of the square grid defining the GTM (total number of nodes $K = \text{NrNodes}^2$). *NrRBF* is the size of the grid locating the radial basis functions (similarly, the total number of functions defining the manifold $M = \text{NrRBF}^2$). RBFw represents RBF width, and RegCoeff is the regularization coefficient

samples of ChEMBL. Any combination of Set2 and Set3 with external compounds Set1 did not lead to better models: the information contained in either Set2 or Set3 is therefore rather representative of ChEMBL. Furthermore, boosting of the frame set by addition of external Set1 structure did not help, but did not harm, either: during the evolution process, top fitness maps were generated with all five frame sets, including Set1 alone. Thus, in agreement with previous observations in context of Kohonen maps [3], it was observed that frame set size is not a key factor.

Furthermore, participation of a compound to the frame set is not a guarantee of being well predicted in GTM-based regression models For instance, absence of inhibitor examples for every second target in Set3 had no impact on the quality of regression-driven $pK_i$ predictions for those targets. With Map 3 based on the Set3, the average cross-validation $pK_i$ RMSE is 0.93 log units over the compounds included in Set3. However, for the remaining compounds, including the ones associated to targets not represented at all in Set3, the corresponding *RMSE* is 0.98. Likewise, for *Set2*-based Map 2, *RMSE* scores for frame and other compounds are of 0.98 and 0.97, respectively.

### *Winning descriptor types*

Pharmacophore- and force-field-colored atom *pair* and atom *triplet* counts were winners (Maps 2, 3, 4 and 5) over more sophisticated sequence and circular fragment counts (latter being represented in Map 1, which however turned out to be less proficient in external validation challenges). They were recurrent in many other suboptimal setup schemes produced, and also tended to show up systematically amongst best performers during preliminary runs serving for technical fine-tuning of the cluster deployment

scheme (results not shown). All these hints suggest that they are intrinsically qualified as consensus descriptor spaces, showing robust neighborhood behavior with respect to many different properties. Winning descriptor spaces are thus:

- relying on chemical context-sensitive atom labeling schemes (force field, pharmacophore type) rather than on the straightforward atom symbol labels. Note that force-field based typing is simply a more detailed atom classification scheme, and is inclusive of pharmacophore typing (pharmacophore types can be rather accurately assigned to any atom of known force field type),
- they systematically belong to the fuzzy, generic subset of the ISIDA fragment descriptor spectrum, defining the least sparse matrices.

### Performance with respect to map selection sets

Figure 4 below illustrates the relative percentages of selection sets successfully passing the cross-validation tests, in both regression and related classification challenges, as a function of the respective success score threshold: $Q^2$ and QBA, respectively.

Although the fitness function used to select these maps was mainly controlled by the mean $Q^2$ value over the 144 selection sets, the left-hand plot looks—at first sight—rather disappointing: none of the maps seems to be able to perform robust quantitative $pK_i$ predictions ($Q^2 > 0.4$) for more than 40 % of the 144 targets. However, if the same selection sets data are analyzed through the less strict prism of active versus inactive discrimination (Challenge 1), the right-hand plot of Fig. 4 depicts a much more positive message: maps are competent with respect to more than 80 % of targets, at $QBA > 0.65$.

### Recognition of target family-specific patterns: discrimination between generic families of binders to a target family

The recognition challenge of ligands/non-ligands of target families—passed with excellent results, as visible from Fig. 5—is a proof that the produced maps are, furthermore, competent to capture the rather fuzzy structural signatures that define ligand families (monoamine and other GPCR, kinase, peptidase, and nuclear receptors).

The monoamine and other-GPCR ligands contrast with respect to homogeneity of the chemical structures. The other-GPCR family was defined by opposition to a well-defined receptor category, thus being more heterogeneous. This is clearly revealed in the Activity Class Maps (ACM) from Fig. 6. Both ACMs were built on hand of the same compound collection—the union of all selection sets—meaning that overall compound densities are the same, as shown by an easy visual check of color density patterns. Within this global collection, the left-hand map highlights the nodes that preferentially harbor monoamine GPCR binders, herein represented by 6000 active ligands. Interestingly, the map does not show any unique monoamine GPCR "activity island" regrouping all the actives within a same area (Fig. 6). Albeit all these ligands rely on the well-

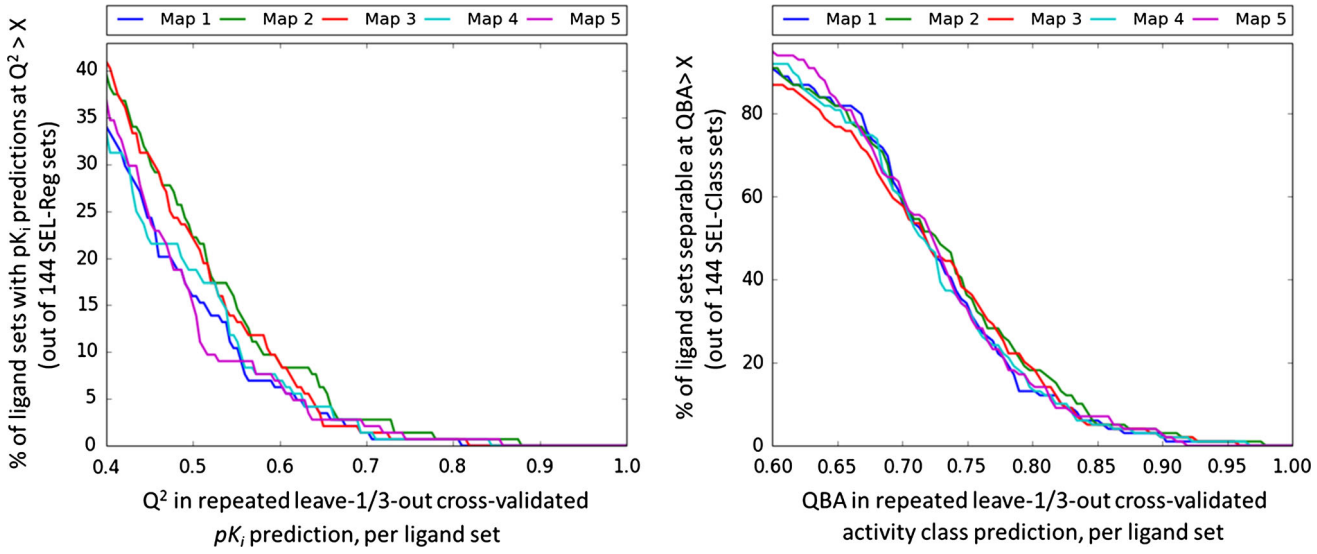



**Fig. 4** Percentages of selection sets successfully passing the cross-validation tests, in both regression and related classification challenges, as a function of the respective success score threshold. Each curve corresponds to one map. The *left plot* represents performances in regression models (selection sets), the right one—classification models (Challenge 1)

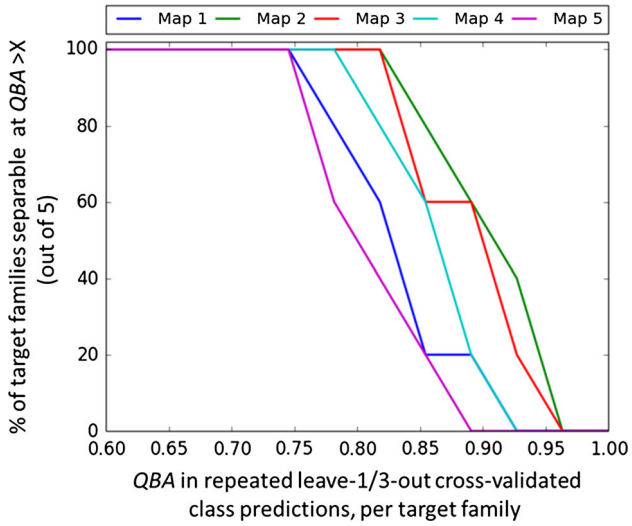



**Fig. 5** Percentages of target families successfully passing the cross-validated classification challenge (Challenge 2), as a function of the balanced accuracy threshold. *Each curve* corresponds to one map

known aromatic-linker-cation pharmacophore, they are not forming, by any standards, a structurally homogeneous family. This common signature (also present in non-member compounds designed as putative actives, but not passing herein adopted activity thresholds) may sometimes represent only a small part of the entire molecule. The successful separation of members from non-members is achieved within many small clusters of structurally homogeneous compounds. Note that this high-resolution break-up of families into local sub-clusters "spontaneously" emerged from the evolutionary process, as the maps were seen to steadily grow in size (the upper limit of the node number had to be manually increased after preliminary trials, results not shown). Also, the higher the resolution, the higher the chance of over-fitting artifacts (by-heart learning) of the method—however, quality scores based on aggressive cross-validation seem to confirm that such high resolution is really necessary.

Other GPCR binders populate distinct CS zones, albeit a partial overlap thereof is expected, and observed (the families actually share promiscuous ligands). The herein rendered 10,633 binders of other-GPCRs are a major source of diversity in medicinal chemistry. Yet, the other-GPCR family is very well separable on the maps, at significant *QBA* values around 0.85 on Maps 2 and 3—best performers of this challenge. This shows that maps are perfectly suited for this type of classification problems, which would make them useful in target-family-focused [50] library design. The presence of nodes in red on both maps indicate regions populated by the members of the other considered classes—kinases, peptidases, nuclear receptors, which occupy quite narrow, well-delimited chemical space areas (see.pdf files in Supporting Information).

## The global ChEMBL activity class recognition challenge

Although the manifold generation is a completely unsupervised learning process, and participation to the frame sets does not enhance the predictivity for concerned molecules, one may still argue that so-far reported results are tainted by the fact that they are based on the same molecules used to select the maps. This section concerns challenges involving an exhaustive set of novel targets, some of which are functionally and structurally completely unrelated to selection sets proteins. The maps are challenged to discriminate between their active and inactive ligands.

Figure 7 reports excellent results with respect to the Challenge 3 (left). Virtually 100 % of these 410 targets succeed in classification, i.e. report *QBA* values above 0.6.

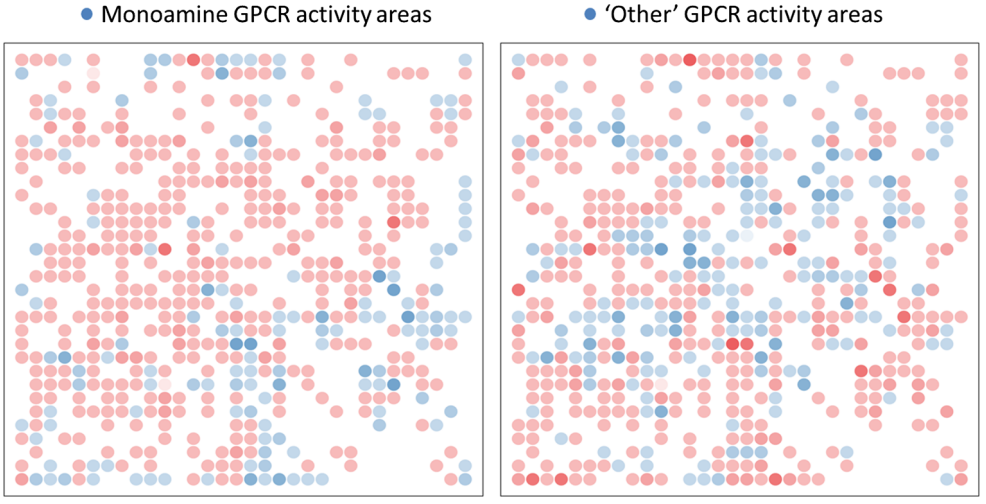



**Fig. 6** Target family-specific 'activity zone' rendering on Map 2. *Blue nodes* are populated mostly by binders, *red* by non-binders to targets within a family. Color intensity scales with the overall population density of a node. *Left map* concerns monoamine GPCRs, right—other GPCRs. The zones populated by monoamine GPCR binders weakly overlap with those corresponding to other GPCRs

Map 2 is providing the top curve of relative best performances.

Results in the Challenge 4 are still satisfactory. However, they are clearly worse than Challenge 3 results, albeit most of challenge 4 sets (392 precisely) are also part of challenge 3. The difference is the very large compound libraries with reported *potency* measures, most likely stemming from noisy, primary high-throughput screening. These sets are (unsurprisingly) recurrently failing the classification challenge.

The GTM model proposed by selected universal maps seems to work "out-of-the-box", without any fitting. The map is merely interpreted on the basis of some labeled instances. Yet, they have an excellent ability to split external ligand sets by activity for never before encountered targets, classes of targets, and even entire organisms. This softer, alternative kind of inductive knowledge transfer is the cornerstone of the universal mapping concept. The GTM formalism not only provides a common referential—a "pharmacophoric latent space"—into which drug-like compounds can be mapped in a NB-compliant manner, but also an intuitive, 2D representation thereof.

For example, Fig. 8 renders the ACMs corresponding to a typical drug discovery target—Cyclooxygenase II, CHEMBL230—and an anti-target—Cytochrome 2C9, CHEMBL3397. It illustrates a typical bias of CS exploration by screening studies. On the right-hand map the 1749 compounds tested on the 2C9 cytochrome cover a wide CS: the red and blue points are scattered rather homogeneously. Tests on cytochromes aim at understanding whether and how drug candidates interact with these key players. On the left-hand side, the 3129 Cox2-associated molecules appear in a more condensed zone of the CS: the red and blue points are less widely spread although the number of compounds they represent is nearly two times larger than the 2C9 dataset. The Cox2-associated compounds are molecules that have been considered interesting enough to deserve the effort of an experimental assessment of Cox2 activity. Thus, the dense regions of the ACM map the chemical space perceived as Cox2-relevant, with the blue nodes outlining the zones predominately populated by actives.

The discovery of Cox2 inhibitors (seen through the prism of ChEMBL data) did not rely much on random screening, but more on stepwise compound optimization. The first selective Cox2 compound had been discovered by DuPont before the actual cloning and characterization of the enzyme, and served as lead for the development of the Coxib series [51]. Compounds tested on this enzyme were primarily molecules already related to anti-inflammatory effects, not random collections.

In this light, the left-hand map in Fig. 8, albeit more focused than the cytochrome counterexample, may not meet the intuitive expectation to see families of known anti-inflammatory compounds cluster neatly on a few nodes. Yet, the map provides a robust separation of Cox-2 actives from inactives (QBA = 0.7).

In fact, the interpretation of ACM is complicated by the fact that GTMs only fuzzily associates a molecule to a node. Representing a molecule as a pattern of responsibilities over the entire set of nodes is unintuitive. It is possible to collapse this information on a single point on the map, as described in previous works [9, 28]. However, this step necessarily triggers some information loss—*visual* separation of actives and inactives pinpointed on the map, is less clear-cut.

For analysis, the Cox-2-associated compounds were loosely assigned to key medicinal chemistry series: coxibs,

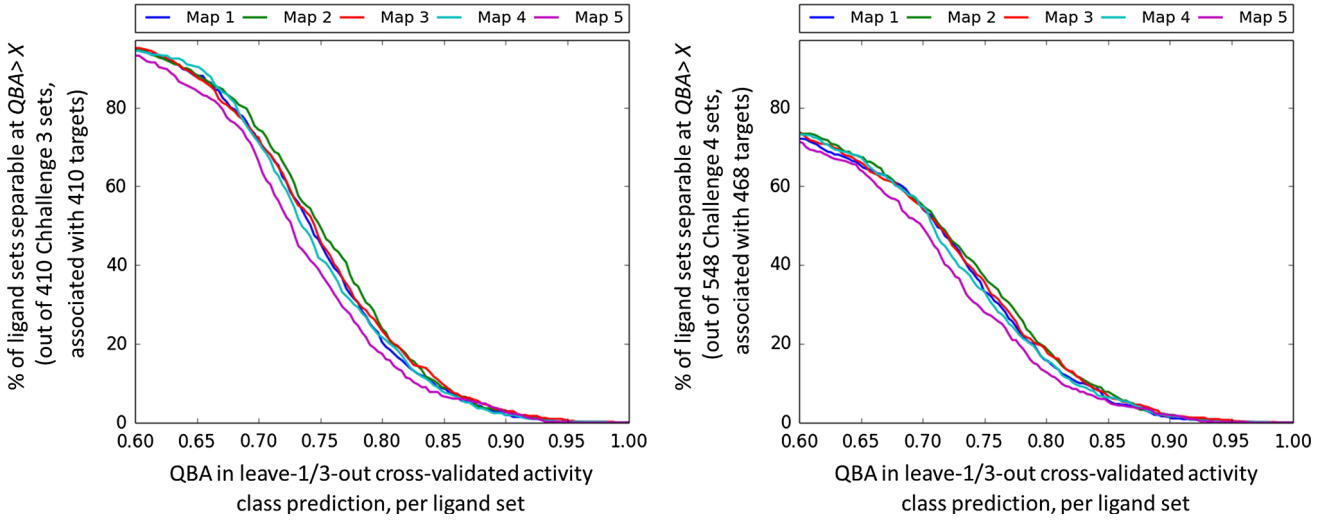



**Fig. 7** Percentages of targets in Challenges 3 (*left*) and 4 (*right*), successfully passing cross-validated activity class recognition tests, as a function of the cross-validated balanced accuracy (QBA) threshold. *Each curve* corresponds to one map

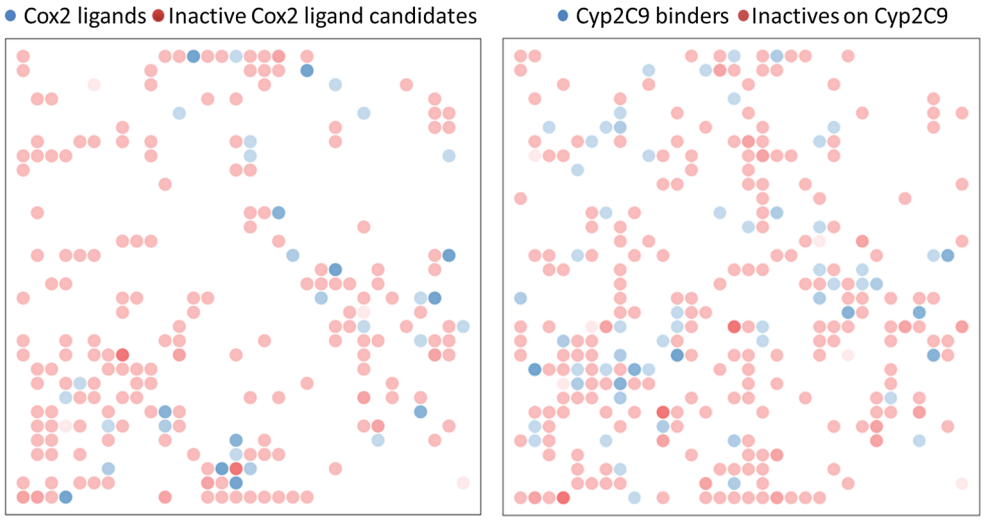



**Fig. 8** ACM rendering on Map 2 for Cox2 ligands (*on the left*) and Cyp2C9 binders (*on the right*). *Blue nodes* are populated mostly by binders, *red* by non-binders to target. While the dataset of Cyp2C9 is almost twice smaller than Cox2, it covers wider CS, as shown by larger spread of colored nodes on the ACM. Cox2 binders, on the other hand, belong to more focused families of compounds

fenamic acids (-fenacs), -profenes, indometacin derivatives, and more recently discovered series of benzopyrans [52] and tetrahydrobenzofurans [53]. The above were the most populated within the employed ChEMBL set, and were assigned by means of ChemAxon substructure searches, using SMARTS [54] definitions for the specific structural signatures of each class. The folder *Cox2-families* in Supporting Information provides the employed SMARTS queries and a text file annotating each of the Cox-2-associated compounds by MedChem family membership. Those not matching either of queries were labeled *other*. The assignment of these families to map nodes is shown in Fig. 9.

For each MedChem family, specific ACMs (not shown) were realized on the basis of family members only. Nodes dominated by actives of each family were marked, in the left-hand part of Fig. 9, with arrows that are color coded by family. Few of pinpointed nodes are red, i.e. not dominated by actives in the family-specific ACMs. These nodes are accommodating more inactives from other classes and therefore dominated by inactives in the global Cox-2 scenario. This is in particular the case with non-specific (and rather weak) Cox-2 binders, which draw their anti-inflammatory effect from interactions with related targets: fenamic acids and the -profene series.

It may also be seen that actives of some families: indomethacin derivatives, and, foremost, benzopyrans, are indeed focused onto one or two nodes. Not mentioning the "Other" category, where diversity is expected, both coxibs—dominant in terms of size—and tetrahydrobenzofuranes, are far from homogeneous, and spread over a significant map area. Thus, initial compound diversity as surprisingly observed in Fig. 9 is, indeed, not only a reflection of multiple chemical series tested on Cox-2, but foremost a consequence of intra-series diversity.

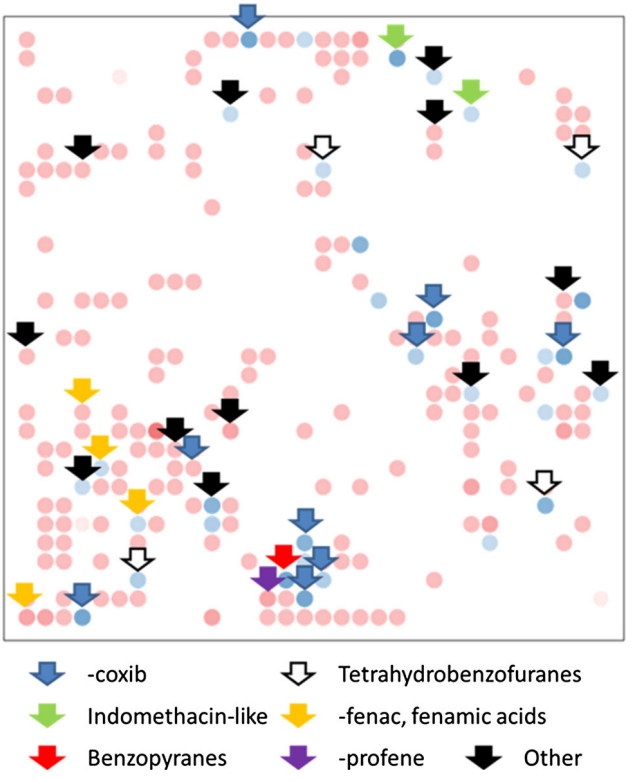



**Fig. 9** Family analysis of the Cox-2 inhibitors. *Blue nodes* are populated mostly by active compounds, *red*—by inactives. Color intensity scales with the overall population density of a node. ACM nodes (of any color) harboring actives, i.e. to which actives of any family contribute mostly, within each MedChem family are marked with *colored arrows*

This, however, is not equivalent to flawed neighborhood behavior. Within the broadly spread coxib-like series, locally homogeneous subfamilies can be clearly localized on the map, as visible in Fig. 10. Mapping does not follow the scaffold-driven paradigm. This is not a surprise, as the underlying topological pair count-based descriptor belongs to a family known for its scaffold hopping ability [55]. Most of the diversity within the broadly defined coxib family stems from the nature of the central ring: plain phenyl, diazoles, oxa/thiazoles, aliphatic hydrophobe or aliphatic functionalized.

Yet, separation is not exclusively driven by the scaffold/ring considerations: substituents are important as well. In particular, ionizable sulfonamides tend to reside in the "northern" areas, while sulfones are most often seen in the southern hemisphere—with one exception: analogues based on aliphatic hydrophobic central rings. In this case, the specific signature of that moiety overrides the $–SO_2NH_2$ versus $–SO_2CH_3$ distinction. Actually (not highlighted on the map, for clarity) sulfones with a central phenyl ring, arguably close analogs of the "northern" sulfonamides are seen to reside in the "southern" aliphatic ring region. With a rather ubiquitous phenyl ring at the core, substituent patterns control the mapping.

While the scaffold-centric perspective is very useful from a scholar point of view, to report and systematize ongoing research, it is intrinsically empirical and not well defined. Such coarse classifications typically regroup compounds of widely varying properties—and many candidates that actually prove to be inactive, as is the case for most of the coxib family members. Furthermore, the biological target per se makes no distinction between scaffold and substituents, but perceives the ligand as a whole. The herein revealed perspective of the map, in conflict with the scholar scaffold-based view, is therefore arguably better. This map was selected to distinguish between actives and inactives of targets different from Cox-2, but appears to successfully solve the problem for this target.

## The drug transporter challenge

This particular challenge was included here because

(a) it focuses on a particular set of macromolecules, evolved to recognize either endogenic compounds (nutrients) or xenobiotics (efflux pumps),
(b) it is based on curated data sets for which independent QSAR studies exists, which creates a good benchmarking opportunity, and
(c) it provides a more thorough analysis of ligand-transporter interactions: specific challenges aimed at discriminating substrates from non-substrates are included in addition to inhibitor/non-inhibitor challenges, similar to already seen ones.

As visible from Fig. 11, these challenges are well managed—in particular by Map 2, which only falls short of separating substrates of BCRP at $QBA > 0.6$ (actual value: 0.58). The least successful models reported concerned MRP4 inhibitors, at $CCR < 0.63$. Therefore, published *fitted* models do not seem to be, overall, much better than the parameter-free

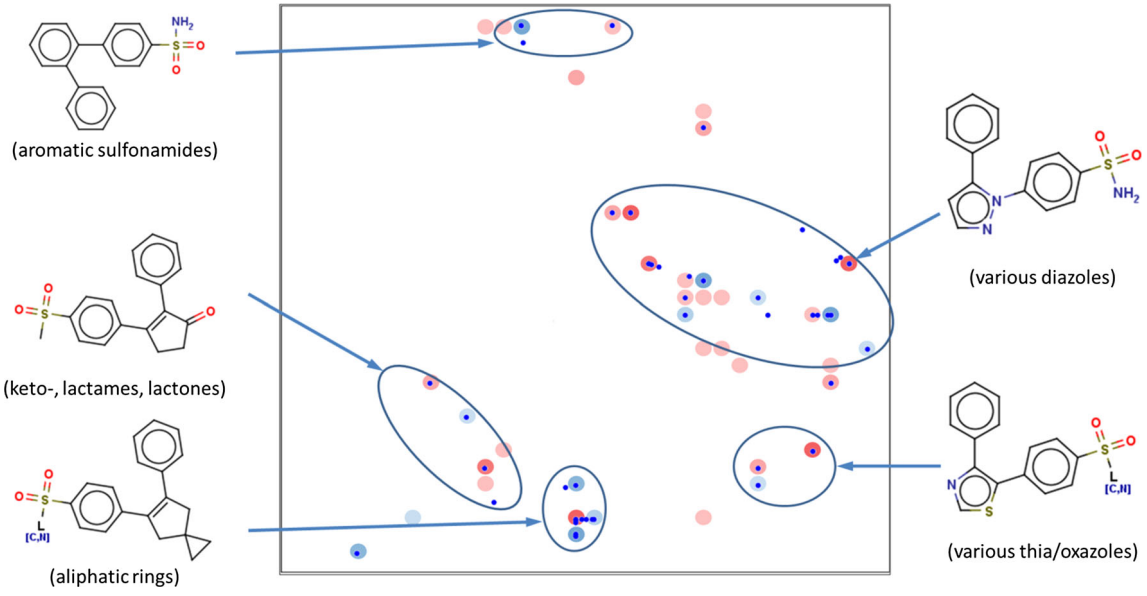



**Fig. 10** Analysis of coxib-like Cox-2 inhibitor family. ACM shows nodes to which coxib-like compounds contribute the most. *Blue dots* demonstrate the mean position of active compounds. Inactive compounds are not shown. Common structural motives of actives found in different regions of the map are demonstrated

extrapolation method provided by GTMs and never confronted to any transport/efflux pump at selection stage.

### In vivo challenges: antiparasitic and antiviral properties

The maps perform surprisingly well in discriminating activity class with respect to complex, whole-organism activities (challenges 6 and 7). Anti-*Plasmodium* activity is an outstandingly complex issue per se, given the very many envisaged antiparasitic action mechanisms, leading to a plethora of distinct testing protocols returning not always concordant results—a same molecule may be active in one protocol, yet inactive according to another. Every protocol was treated as stand-alone binary class property to predict. The antimalarial challenge is technically similar to the global ChEMBL and transporter activity classification challenges, except that now the "targets" are the various anti-*Plasmodium* testing protocols. The 17 herein selected sets may reach sizes of up to 1000 molecules. As can be seen from Fig. 12, GTM models perform well for a vast majority of classes—Map 4 manages to separate all properties at $QBA > 0.6$, but Map 2 keeps the upper hand in terms of very well separated problems.

Antiviral family classification is run, similarly to target family classification, against the common background of all ChEMBL compounds found to display some antiviral activity, all classes confounded. First, Fig. 13(left) shows that the ~30,000 antivirals do cover a significant area of the drug-like space, in which compound sets associated to each virus family can be very well separated by the Maps 2 and 3, in particular.

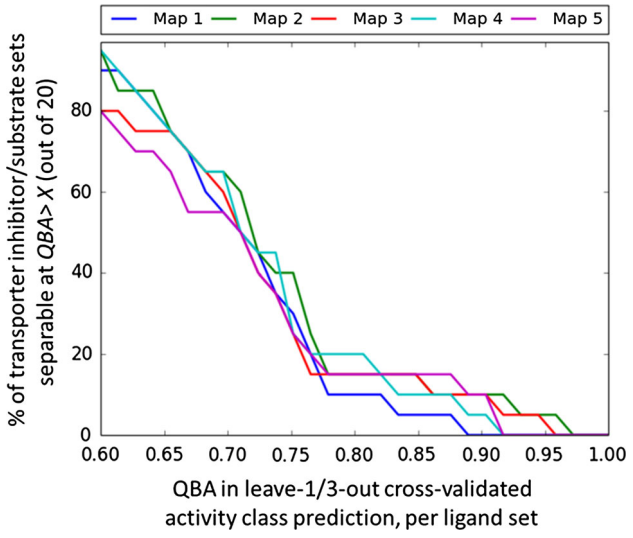



**Fig. 11** Percentages of challenge 5 sets successfully passing cross-validated classification tests, as a function of the cross-validated balanced accuracy (QBA) threshold. *Each curve* corresponds to one map

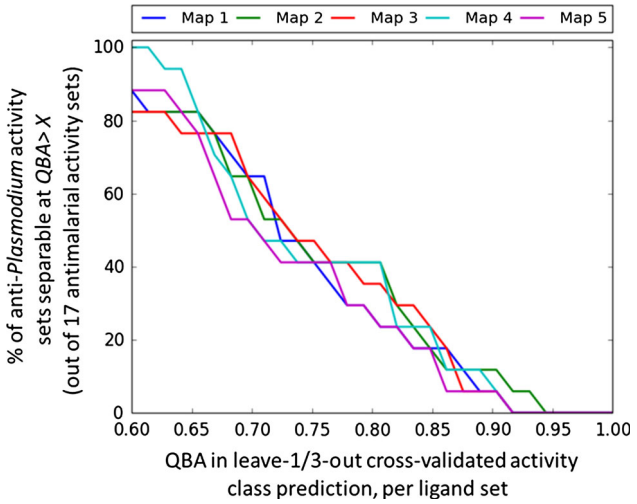



**Fig. 12** Percentages of challenge 6 sets successfully passing cross-validated classification tests, as a function of the cross-validated balanced accuracy (QBA) threshold. *Each curve* corresponds to one map

With all maps, the least obvious classification challenge is the recognition of anti-lentivirus compounds. However, even this problem is honorably solved by Maps 2 ($QBA = 0.68$) and 3 ($QBA = 0.70$).

### Is chemical space coverage of related targets similar?

Table 3 records the definition of each family, sorted by decreasing $p$ value. For instance, the cohesion score of 0.44 signals the existence of a common chemical subspace associated to Adenosine receptors. It is distinct from the CS associated to other GPCRs, with a separation value of 0.93. Indeed, many specific subfamilies of GPCRs are nowadays well characterized, and for them the structural signatures of associated ligands are rather well known. Most adenosine receptors inhibitors [56, 57] are, in fact, derivatives of adenosine, theophylline, or xanthine, and are selective binders. Clearly, the "ligand view" of target similarity—in this and all the other publications exploring this topic—is at risk of being biased because based on only so-far known ligands, themselves often generated in virtue of the similarity principle. This reluctance to leave well-explored chemical space zones is a rational strategy only if, indeed, they happen to be the only compatible with the given receptor. Therefore, the herein built maps (like all so-far published chemoinformatics models) are a consistent display of the current knowledge status, and not a first-principle approach predicting all the possible binding modes to a target. They were built and tested on an extensive amount of binding data. Note that all so-far known ligands retrieved from ChEMBL equally contribute to cumulative responsibility vectors defining the CS space coverage of a

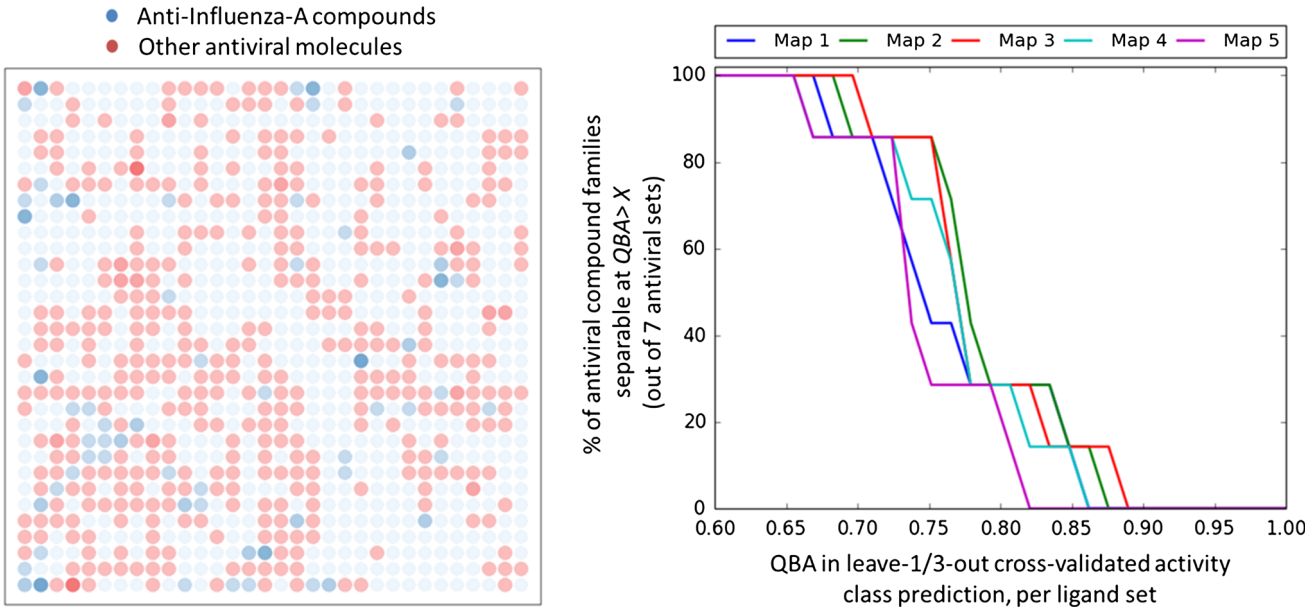



**Fig. 13** *On the left* ACM for anti-Influenza-A compounds. *On the right* percentages of challenge 7 sets successfully passing cross-validated antiviral activity classification tests, as a function of the cross-validated balanced accuracy (QBA) threshold. *Each curve* corresponds to one map

**Table 3** Statistical analysis of distances between the CS zones associated to various GPCRs and kinases, classified into subfamilies

| Super-family | Family | Family size | Shortest inter-target distance In family | Cohesion and SD In family | Shortest inter-target distance To others | Separation and SD To others | *p* value |
|---|---|---|---|---|---|---|---|
| gpcr | Adenosine | 4 | 0.182 | 0.442 ± 0.165 | 0.936 | 0.993 ± 0.011 | 1.00E−09 |
| Kin | TK | 35 | 0.035 | 0.408 ± 0.148 | 0.043 | 0.490 ± 0.237 | 1.00E−09 |
| gpcr | Serotonin | 8 | 0.38 | 0.816 ± 0.164 | 0.54 | 0.977 ± 0.050 | 6.00E−07 |
| gpcr | Opioid | 4 | 0.202 | 0.481 ± 0.263 | 0.917 | 0.991 ± 0.013 | 2.40E−05 |
| gpcr | Melanocortin | 4 | 0.322 | 0.644 ± 0.186 | 0.813 | 0.989 ± 0.022 | 5.60E−05 |
| gpcr | Prostanoid | 8 | 0.099 | 0.823 ± 0.203 | 0.472 | 0.981 ± 0.046 | 7.40E−05 |
| gpcr | Dopamine | 5 | 0.253 | 0.688 ± 0.266 | 0.54 | 0.972 ± 0.056 | 0.0021 |
| gpcr | EDG | 4 | 0.378 | 0.713 ± 0.195 | 0.719 | 0.964 ± 0.053 | 0.0062 |
| gpcr | Nucleotide-like | 6 | 0.182 | 0.774 ± 0.290 | 0.472 | 0.990 ± 0.033 | 0.0074 |
| gpcr | Somato-statin | 4 | 0.115 | 0.606 ± 0.308 | 0.502 | 0.977 ± 0.045 | 0.0086 |
| gpcr | Adrenergic | 7 | 0.035 | 0.798 ± 0.314 | 0.804 | 0.984 ± 0.023 | 0.0091 |
| gpcr | Histamine | 4 | 0.644 | 0.863 ± 0.112 | 0.756 | 0.965 ± 0.046 | 0.042 |
| Kin | CMGC | 8 | 0.171 | 0.401 ± 0.138 | 0.067 | 0.448 ± 0.216 | 0.094 |
| Kin | AGC | 12 | 0.102 | 0.570 ± 0.295 | 0.036 | 0.525 ± 0.272 | 0.23 |
| Kin | Src | 6 | 0.17 | 0.412 ± 0.156 | 0.035 | 0.454 ± 0.217 | 0.33 |
| Kin | CAMK | 9 | 0.053 | 0.438 ± 0.345 | 0.034 | 0.469 ± 0.286 | 0.6 |

Column 3 reports how many receptor subtypes are counted within each family—the Adenosine family counts 4 receptors, etc. The explicit listing thereof is given as Supporting Information. "In family" columns refer to distances between target pairs that are both members of the family (cohesion and shortest intra-family distance). "To others" concern distances between a family member and a member of any other family (separation and shortest distances within each super-family). The *p* value (Student's test) compares "in family" and "to others" distance distributions

target: this is not the expression of some specific, closed medicinal chemistry series of analogues (unless the one specific series is the only knowledge one has with respect to that target). So-far unknown ligands do not contribute, obviously. In spite of all the limitations and biases, *p* values below are, more often than not, clearly showing that CS coverage supports biological classification—intrinsically validating the universal character of the maps.

Histamine receptors allow for much more diversity in terms of binders than Adenosine receptors, and yet clearly have something in common: even though the ligand space of one histamine receptor does not strongly resemble the one covered by another, it still resembles more than any other CS zone addressed by non-histaminic receptors. Then, not all members of a SARFARI family must fit equally well into the family—this study only focuses on mean CS coverage as resulting from unbiased cumulative responsibility vectors. In practice, one may want to have a closer look at the internal family cohesion, and use this information in focused library design. The more cohesion (i.e. the smaller the average intra-family distance), the safer it is to assume that a focused library, composed of ligands with responsibility vectors similar to the family signature, will be promiscuous and hit all the family members.

Sometimes, cohesion can be only marginally better than separation. For example, the CS associated to a Tyrosine Kinase (TK) is only marginally denser (cohesion value 0.407) than the CS of non-TK kinases (separation value 0.490). The absolute shift is actually much lower than the cumulated standard deviations of the two means: the low *p* value suggests that TKs form a more homogeneous subset of all kinases. Remarkably, the two closest TK targets have a nearly perfect CS overlap (distance 0.035). However, the closest non-TK target is located at a comparable distance 0.043 from a TK member. In general, kinases are much more focused on a common, specific CS—within which it is very difficult to discover selectivity zones. All these observations perfectly match common medicinal chemistry observations: some TK-specific families of compounds (tyrphostins, lavendustins, quinazolines) do exist [58–60], all while other series of compounds are highly promiscuous [61, 62]. The fact that these observations could be 'read' from the map is an additional proof of its relevance.

## Conclusions

This work attempted to address the question whether a Universal, compound set-independent Generative Topographic Map can be generated. The claim of universality is quantitatively justified, with respect to all the structure–activity information available so far. To this purpose, an evolutionary map growth and selection procedure considered both the choice of meta-parameters (frame molecule sets, descriptor types) and map-specific parameters (size, RBF function controls, etc.) as degrees of freedom. It was associated to a fitness function measuring the polypharmacological NB-compliance of the map, with respect to a multi-target quantitative affinity prediction challenge. The optimal manifolds were seen to grow in rather low-resolution molecular descriptor spaces: pharmacophore- or force-field-type colored atom pairs and triplets rather than more specific sequence or circular fragment counts included in the pool of competing ISIDA descriptor types.

The herein investigated machine learning paradigm is, in our opinion, quite novel because it does not match either of two dominating approaches used in chemoinformatics for polypharmacological prediction: classical multi-target QSAR and chemogenomics. Like in classical QSAR, the approach generates one individual model per target. It does not require, like in chemogenomics, target-specific descriptors. However, unlike classical QSAR, which requires tedious fitting of each individual model, the key advantage of our approach is that all tunable parameters and meta-parameters (descriptor choice) are already determined, and were shown to represent a good choice for a plethora of classification models of biological properties (target binding and in vivo activities) which were completely unrelated to properties used for fitting. In other words, models are still target-specific (one needs to project, for each property to model, a dedicated structure–property data set onto the map), but their parameters are not. These parameters are transferable not only to targets similar to the ones used for training—as expectable on the basis of the principle of chemogenomics—but to a vast majority of arbitrary biological properties complying to the similarity principle "Similar ligands have similar activities". This transferability may indeed be seen as a form of inductive transfer of knowledge, a major source of enhancement in chemogenomics [32]. Note, however, that the herein witnessed transfer is not of the same nature as in chemogenomics computations, where rich knowledge with respect to well-studied targets is used in order to support building a predictive model for a *related* but little studied (or outright orphan) target with insufficient (or inexistent) experimental structure–activity data. Here, the transfer does, on one hand, not seem to be conditioned by the relatedness between targets, but, on the other, only concerns the model building recipe (descriptor choice and manifold geometry) and not target-specific information. Therefore, in its present implementation, the universal maps may prove to simplify modeling of activities of little studied targets (for which available structure–activity data could not support independent model fitting and validation) but are not effective in target deorphanization.

One may thus think about the approach as the most general (the most generally applicable, property-independent) computational embodiment of the similarity principle so far reported, based on a non-linear transformation of a best-suited initial descriptor set. The ability to generate "plug-and-play" models for arbitrary targets, with zero fitting effort (just add reference data) but a considerable chance of success represents, per se, an important and unexpected methodological progress—not mentioning the further putative benefits, such as visualization.

The GTM manifold is of unsupervised nature: it does not contain any target-related structure–activity information. Yet it can be colored by property, by mapping of a dataset with associated experimental labels. New instances are annotated according to the colored manifold (ACM, or Activity landscape). This procedure was at the core of proving the Universal character of evolved maps. The maps are perfectly suited to solve classification problems concerning chemical structures never used to fit the map: on the overall, more than 80 % of the more than 600 distinct and varied classification problems, chosen such as to cover a maximum of exploitable SAR data, were successfully solved. This justifies, in our view, the claim of "universality" of the constructed GTMs.

Nevertheless, the present article is not about optimal predictive power for individual models. Supervised learning based on property-specific descriptor spaces may, rather than relying upon a common consensus map based on a consensus descriptor space, be the best strategy in that respect. This is unless the common GTM frame could facilitate inductive transfer of knowledge between targets (going beyond above-mentioned transferability of parameters, and thus acquiring the advantages of multi-task learning and enable target deorphanization approaches)—an interesting aspect that will be the topic of further work. The non-trivial news so far is that such consensus maps can be found, and that in spite of the compromise setup, which is probably not optimal for any single property prediction, the predictive power remains reasonable for a vast majority of properties. In perspective, it would be interesting to compare each GTM-based model to an optimal, dedicated approach in order to assess how much predictive power is lost or gained, if ever above-invoked inductive transfer is shown to occur.

In addition, the maps provide with an intuitive representations of the data. They were shown to provide a consistent analysis of the drug-like space. Mapping clearly does not follow the popular scaffold-centric view. However, the mapping approach is arguably meaningful because the estimated activities are generally correct. Due to quantitative validation, the user may gain confidence in the rendered visual patterns, and draw very meaningful conclusions on their behalf.

These GTM models will, in our opinion, be quite helpful to provide a better global overview of modern trends and challenges in medicinal chemistry. So far, the accent was set on highlighting the fact that the approach fulfills classical expectations from a polypharmacological prediction tool, namely that (a) positioning of ligand candidates on the map successfully results in prediction of their biological property profiles, and (b) targets may be meaningfully characterized by the signatures of cumulated responsibilities of their ligands, which are consistent with the accepted classification of targets in families. Whilst this article strived to reveal the expected, known inter-target relationships in order to underline the robustness of the approach, significant overlap of a priori uncorrelated targets may signal unexpected drug repositioning opportunities.

## Supporting information

The zip file *SuppInfo.zip* unpacks as a directory 'SuppInfo'. Files therein were mentioned and described in the relevant paragraphs of the publications. This material is available free of charge via the Internet.

**Acknowledgments** The Laboratory of Chemoinformatics wishes to thank the High Performance Computing centers of the University of Strasbourg, France and the Babes-Bolyai University of Cluj, Romania for supplied computer power, and assistance. Many thanks to Prof. Jürgen Bajorath, for providing the clean and coherent ChEMBL compound subset. K. Klimenko, B. Viira and T. Gimadiev are acknowledged for the help with preparation of antiviral, antimalarial and transporters datasets. PS and AV thank Russian Scientific Foundation (Agreement No 14-43-00024 of October 1, 2014) for support.

## Appendix: ChEMBL data curation protocol

This "Appendix" describes the data curation protocol concerning Challenges 3 and 4, as shown in Fig. 14 and explained in further details below.

First, the complete list of biological targets of *Homo sapiens* was retrieved from ChEMBL. For each target labeled as "single protein" (2474 proteins) associated ligand activity data were uploaded using a script available as Supporting Information. Analysis and binning of activity data was done as follows:

1. Compounds associated with an inhibition percentage equal or less than 50 % are labeled as inactives.
2. Compounds with an associated dose-dependent activity value lower than an imposed activity threshold are labeled as actives.
3. Compounds with an associated dose-dependent activity value equal or higher than ten times the activity threshold were labeled as inactives.

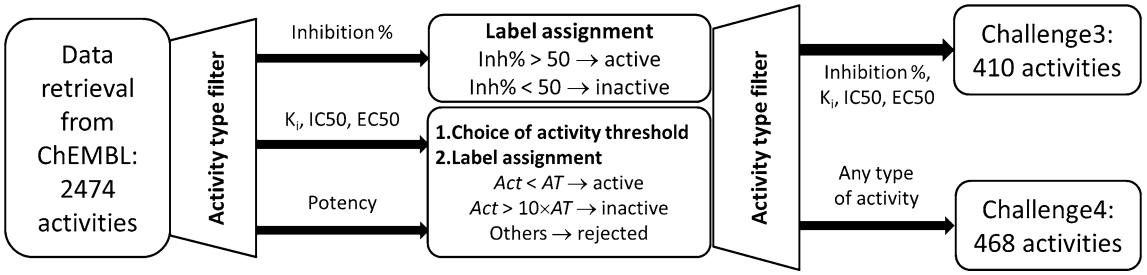



**Fig. 14** Workflow of data mining and curation for external *Homo sapiens*-target related validation. For mined targets, compounds were labeled as active or inactive according to their reported activity type: inhibition %, $K_i$, IC50, EC50, "potency"—and a chosen activity threshold (*AT*). Targets that have only potency activity type associated entered only challenge 4, others entered both challenges

Eventually, compounds with multiple entries leading to contradictory tentative activity class assignments were ignored. Compounds for which label active or inactive could not be set were also ignored.

In the steps 2 and 3, several activity thresholds were tested: 1000, 500, 100 and 50 nM. A threshold was used for a given target if:

- more than 100 compounds could be successfully classified for the given target, and
- out of these, at least 20 compounds are active, and
- inactives are representing more than 50 % of the set.

If the above rules could not be satisfied, the target was discarded. If more than one threshold value satisfied these conditions, the resulting dataset with the proportion of active compounds closest to 25 % was kept.

If the activity value used in the previous steps was the $K_i$, IC50 or EC50, the targets entered the both Challenge 3 and Challenge 4. If the activity type was *potency*, the associated target entered the Challenge 4 only. Those data are results of high-throughput screening (HTS) campaigns.

Supporting Information features two tar archives, *Challenge 3.tar.gz* and *Challenge 4.tar.gz* respectively, containing files named *<ChEMBLTarget_ID>_S<subset number>.class* which report, for the named target, a list of ligand ChEMBL IDs next to their attributed class (1-inactive/2-active). Distributions contain also the files *target.chid_AT_name*, listing on three columns the validated target ChEMBL IDs, associated activity thresholds AT (in nM) and the name of the target as given in ChEMBL database.

## References


1. Virshup AM, Contreras-Garcia J, Wipf P, Yang W, Beratan DN (2013) J Am Chem Soc 135(19):7296
2. Reker D, Rodrigues T, Schneider P, Schneider G (2014) PNAS 111(11):4067
3. Bonachera F, Marcou G, Kireeva N, Varnek A, Horvath D (2012) Bioorg Med Chem 20:5396
4. Kohonen T (2001) Self-organizing maps. Springer, Heidelberg
5. Agrafiotis DK, Rassokhin DN, Lobanov VS (2001) J Comput Chem 22(5):488
6. Agrafiotis DK (2003) J Comput Chem 24(10):1215
7. Sander T, Freyss J, von Korff M, Rufener C (2014) J Chem Inf Model 55(2):460
8. Gaspar H, Marcou G, Horvath D, Arault A, Lozano S, Vayer P, Varnek A (2013) J Chem Inf Model 53(12):3318
9. Kireeva N, Baskin II, Gaspar HA, Horvath D, Marcou G, Varnek A (2012) Mol Inf 31(3–4):301
10. Oprea TI, Gottfries J (2001) J Comb Chem 3(2):157
11. Renner S, van Otterlo WAL, Dominguez Seoane M, Mocklinghoff S, Hofmann B, Wetzel S, Schuffenhauer A, Ertl P, Oprea TI, Steinhilber D, Brunsveld L, Rauh D, Waldmann H (2009) Nat Chem Biol 5(8):585
12. Lloyd DG, Golfis G, Knox AJS, Fayne D, Meegan MJ, Oprea TI (2006) Drug Discov Today 11(3–4):149
13. Matero S, Lahtela-Kakkonen M, Korhonen O, Ketolainen J, Lappalainen R, Poso A (2006) Chemom Intell Lab Syst 84(1–2):134
14. Öberg T, Iqbal MS (2012) Chemosphere 87(8):975
15. Kauvar LM, Villar HO, Sportsman JR, Higgins DL, Schmidt DE Jr (1998) J Chromatogr B Biomed Sci Appl 715(1):93
16. Horvath D, Lisurek M, Rupp B, Kühne R, Specker E, von Kries J, Rognan D, Andersson CD, Almqvist F, Elofsson M, Enqvist P-A, Gustavsson A-L, Remez N, Mestres J, Marcou G, Varnek A, Hibert M, Quintana J, Frank R (2014) ChemMedChem 9(10):2309
17. Abad-Zapatero C, Perišić O, Wass J, Bento AP, Overington J, Al-Lazikani B, Johnson ME (2010) Drug Discov Today 15(19–20):804
18. Paolini GV, Shapland RHB, van Hoorn WP, Mason JS, Hopkins AL (2006) Nat Biotech 24(7):805
19. Ruddigkeit L, van Deursen R, Blum LC, Reymond J-L (2012) J Chem Inf Model 52(11):2864
20. Reymond J-L, Ruddigkeit L, Blum L, van Deursen R (2012) Wiley Interdiscip Rev Comput Mol Sci 2(5):717
21. Polishchuk PG, Madzhidov TI, Varnek A (2013) J Comput Aided Mol Des 27(8):675
22. Horvath D, Koch C, Schneider G, Marcou G, Varnek A (2011) J Comput Aided Mol Des 25(3):237
23. Horvath D, Jeandenans C (2003) J Chem Inf Comput Sci 43:691
24. Ruggiu F, Marcou G, Varnek A, Horvath D (2010) Mol Inform 29(12):855
25. Varnek A, Fourches D, Horvath D, Klimchuk O, Gaudin C, Vayer P, Solov'ev V, Hoonakker F, Tetko Iv, Marcou G (2008) Curr Comput-Aided Drug Des 4(3):191
26. Varnek A, Fourches D, Solov'ev V, Klimchuk O, Ouadi A, Billard I (2007) Solv Extr Ion Exch 25(4):433
27. Gaspar HA, Baskin II, Marcou G, Horvath D, Varnek A (2015) Mol Inform. doi:10.1002/minf.201400153

28. Gaspar HA, Baskin II, Marcou G, Horvath D, Varnek A (2014) J Chem Inf Model 55(1):84
29. Bishop CM, Svensén M, Williams CK (1998) Neural Comput 10(1):215
30. Horvath D, Brown J, Marcou G, Varnek A (2014) Challenges 5(2):450
31. Bieler M, Heilker R, Koeppen H, Schneider G (2011) J Chem Inf Model 51(8):1897
32. Brown JB, Okuno Y, Marcou G, Varnek A, Horvath D (2014) J Comput Aided Mol Des 28(6):597
33. Lin H, Sassano MF, Roth BL, Shoichet BK (2013) Nat Methods 10(2):140
34. Keiser MJ, Roth BL, Armbruster BN, Ernsberger P, Irwin JJ, Shoichet BK (2007) Nat Biotech 25(2):197
35. Horvath D, Marcou G, Varnek A (2013) J Chem Inf Model 53(7):1543
36. ChemAxon (2009) Standardizer http://www.chemaxon.com/jchem/doc/user/standardizer.html. Accessed Feb 2008, Budapest
37. ChemAxon (2007) pKa calculator plugin https://www.chemaxon.com/products/calculator-plugins/property-predictors/. Accessed Feb 2013. ChemAxon, Budapest
38. Varnek A, Fourches D, Hoonakker F, Solov'ev VP (2005) J Comput Aided Mol Des 19(9–10):693
39. Laboratoire de Chemoinformatique Strasbourg (2012) Nomenclature of ISIDA fragments
40. Gaulton A, Bellis LJ, Bento AP, Chambers J, Davies M, Hersey A, Light Y, McGlinchey S, Michalovich D, Al-Lazikani B, Overington JP (2011) Nucl Acids Res 40(D1):D1100
41. Irwin JJ, Shoichet BK (2005) J Chem Inf Model 45(1):177
42. Sedykh A, Fourches D, Duan J, Hucke O, Garneau M, Zhu H, Bonneau P, Tropsha A (2013) Pharm Res 30(4):996
43. Elhabiri M, Sidorov P, Cesar-Rodo E, Marcou G, Lanfranchi DA, Davioud-Charvet E, Horvath D, Varnek A (2015) Chem A Eur J 21–8:3415
44. Lanfranchi DA, Cesar-Rodo E, Bertrand B, Huang H-H, Day L, Johann L, Elhabiri M, Becker K, Williams DL, Davioud-Charvet E (2012) Org Biomol Chem 10(31):6375
45. Muller T, Johann L, Jannack B, Bruckner M, Lanfranchi DA, Bauer H, Sanchez C, Yardley V, Deregnaucourt C, Schrevel J, Lanzer M, Schirmer RH, Davioud-Charvet E (2011) J Am Chem Soc 133(30):11557
46. Davioud-Charvet E, Delarue S, Biot C, Schwöbel B, Boehme CC, Mössigbrodt A, Maes L, Sergheraert C, Grellier P, Schirmer RH, Becker K (2001) J Med Chem 44(24):4268
47. Elhabiri M, Sidorov P, Cesar-Rodo E, Marcou G, Lanfranchi DA, Davioud-Charvet E, Horvath D, Varnek A (2015) Chem A Eur J. doi:10.1002/chem.201403703
48. Willett P, Barnard JM, Downs GM (1998) J Chem Inf Model 38:983
49. Welch BL (1947) Biometrika 34:28
50. Rolland C, Gozalbes R, Nicolai E, Paugam MF, Coussy L, Barbosa F, Horvath D, Revah F (2005) J Med Chem 48:6563
51. Flower RJ (2003) Nat Rev Drug Discov 2(3):179
52. Wang JL, Aston K, Limburg D, Ludwig C, Hallinan AE, Koszyk F, Hamper B, Brown D, Graneto M, Talley J, Maziasz T, Masferrer J, Carter J (2010) Bioorg Med Chem Lett 20(23):7164
53. Janusz JM, Young PA, Ridgeway JM, Scherz MW, Enzweiler K, Wu LI, Gan L, Chen J, Kellstein DE, Green SA, Tulich JL, Rosario-Jansen T, Magrisso IJ, Wehmeyer KR, Kuhlenbeck DL, Eichhold TH, Dobson RLM (1998) J Med Chem 41(18):3515
54. DayLight (2007) SMARTS http://www.daylight.com/dayhtml/doc/theory.smarts.html. Accessed Oct 2014. Daylight Chemical Information Systems
55. Schneider G, Schneider P, Renner S (2006) QSAR Comb Sci 25:1162
56. Jacobson KA, Van Galen PJM, Williams M (1992) J Med Chem 35(3):407
57. Poulsen S-A, Quinn RJ (1998) Bioorg Med Chem 6(6):619
58. Groundwater PW, Solomons KRH, Drewe JA, Munawar MA (1996) Protein tyrosine kinase inhibitors. In: Ellis GP, Luscombe DK (eds) Progress in medicinal chemistry, vol 33. Elsevier, Amsterdam, p 233
59. Lawrence DS, Niu J (1998) Pharmacol Ther 77(2):81
60. Levitzki A (1999) Pharmacol Ther 82(2–3):231
61. Davies S, Reddy H, Caivano M, Cohen P (2000) Biochem J 351:95
62. Bain J, Plater L, Elliott M, Shpiro N, Hastie C, Mclauchlan H, Klevernic I, Arthur J, Alessi D, Cohen P (2007) Biochem J 408:297

# Chapter 6

# Chemical space of antimalarial compounds

The centerpiece of this thesis is the analysis of the chemical space. The chemical space, defined by the dataset of relevant compounds (encoded in molecular descriptors) and the similarity metrics between them, is not *per se* perceivable by a human. Dimensionality reduction and visualization techniques are widely used to enable such an analysis, but again, in case of thousands of molecules (and in our case, there are thousands of compounds in our antimalarial space), the visualization alone doesn't always help. The Generative Topographic Mapping calculates the probability distribution function for every molecule in the latent space in form of a responsibility vector. The responsibility allows to complement the visual analysis with a specific numerical approach - responsibility pattern analysis.

In this chapter, we describe the application of the GTM, with the novel concepts of universal maps and responsibility patterns, to visualize and analyze the chemical space of antimalarial compounds, and discuss the results and conclusion that can be drawn from such an analysis.

## 6.1 Maps of antimalarial chemical space

Three kinds of maps are considered in this part of the project. The first one is the universal maps described in the previous section. The second type are the global maps of antimalarial compounds, which are built on the MalariaDB and encompass the totality of data collected in this work. Lastly, a local map of antimalarials is selected to maximize the classification performance on ActivityDB data. The optimization parameters,

frame and color sets used for the construction of each type of maps are explained in this section.

### 6.1.1 Universal maps of drug-like compounds

The frame sets and optimized functions are explained in the previous chapter. While the universal maps are not directly connected to any antimalarial activity, they are claimed to be universal for drug-like chemical space and, thus, relevant for antimalarial drugs, which has been shown by their good performance in classification tasks for ActivityDB. For the parameters of the universal maps refer to the Table 5.2.

### 6.1.2 Global maps of antimalarial compounds

These maps were based on the MalariaDB. The meta-parameters are chosen as following:

- Only one working hypothesis for the frame set is considered in this case – the totality of data in the MalariaDB. Since all compounds of this dataset have been tested in an antimalarial bioassay, they should be sufficient to span the relevant antimalarial chemical space. The frame set thus contains 15000 compounds. This set is big enough to provide good chemical diversity, but not too big for the calculations to become too costly.
- The optimized function is the performance of the GTM model in classification of compounds by their mode of action. The classes are assigned following the target grouping strategy described in chapter 4. One-versus-all binary classification is considered: among 1140 compounds in the color set, those that belong to a certain target group or pathway are labeled “active”, and others are ‘ìnactive”.

After the optimization, four top maps have been selected. All of them have demonstrated good performance in all 8 classification tasks. Additionally, they have been built on different descriptor types, and thus correspond to different initial spaces. The parameters of the maps and their scores are shown in Table 6.1.

As the results show, the most relevant chemical space is given by the circular fragments, since they describe most information about relevant chemical environment of all atoms in a molecule. That is consistent with the optimized score: the distinction between the mechanisms is mostly scaffold-driven, and different chemotypes would correspond to different modes of action, while variance in substituents would lead to the modification of measured activity value rather than change the mechanism.

| Map | Descriptors | Resolution | Score |
|---|---|---|---|
| 1 | IIAB-FC-1-3, Atom-centered fragments, atoms and bonds length 1 to 3, with formal charges | $23 \times 23$ | 0.844 |
| 2 | IIA-PH-1-2, pharmacophore-colored atom-centered fragments, with information of atoms only, of length 1 to 2 | $26 \times 26$ | 0.822 |
| 3 | IIAB-1-3, Atom-centered fragments with information on atoms and bonds, of length 1 to 3 | $38 \times 38$ | 0.825 |
| 4 | IAB-FF-P-2-6: force-field type colored counts of atom pairs, 1 to 5 bonds apart, including information on bonds nearest to terminal atom | $40 \times 40$ | 0.828 |

TABLE 6.1: Descriptors types and map resolutions for top global antimalarial maps. For descriptors, ISIDA nomenclature is given. The score indicated here is the average BA of 8 classification tasks.

### 6.1.3 Local map of antimalarial compounds

One map has been selected as the optimal local map of the antimalarial chemical space. The map is named "local" because it has been optimized for the maximum active/inactive classification of compounds is ActivityDB, which is a small portion of MalariaDB. The meta-parameters that are selected for the optimization for this map are the following:

- Four different frame sets are considered as working hypotheses for manifold fitting. Two of these, labeled FS1 and FS2, are (disjoined) random picks from the global set of 250K Malaria-associated ChEMBL compounds, of 10K compounds each. No further checking of the activity flag or conditions of measurement of their activity was undertaken. They are thought to represent the chemical space considered by experts to be relevant for antimalarial research – i.e. compounds that are worth testing against *Plasmodium*. Alternatively, two additional working hypotheses labeled FS1P and FS2P respectively were obtained by adding, to FS1 and FS2 respectively, 1700 random activity-annotated compounds from ChEMBL series with dose-response data (pIC50, pEC50), in order to ensure that large high-throughput screening libraries tested at a single concentration do not completely control the frame sets.

- The end function of the manifold optimization is the best separation of molecules that are active against *Plasmodium* from inactive ones. Thus the color sets are the 17 antimalarial activity measures from ActivityDB. The goal of this map is to provide good classification performance rather than effective span of antimalarial chemical space, since the structures from ActivityDB are not very chemically diverse.

After the GTM parameter optimization, one best map has been chosen. The size of the map is 49×49 nodes, and the descriptor type selected by the genetic algorithm as the best initial space is atom-centered fragments of length 1 to 2, with information on both atoms and bonds, and colored by force-field types. Thus, fragments retain most information about relevant chemical environment of all atoms in a molecule. The map has shown good performance in 3-fold cross-validated classification task for all 17 key data sets, with average balanced accuracy score of 0.80. Distinction between active and inactive compounds is more difficult (resulting in lower BA) for diverse sets of *in vivo* cytotoxicity studies.

The optimal frame set was FS2P, featuring the additional 1.7K compounds with dose-response activity data. Yet, only 42 of the latter are present within the 17 key compound sets. This negligible overlap is not an important factor, since (a) the overlap is small and (b) maps of near-equal statistical robustness (results not shown) could be obtained on the basis of all four frame set choices (FS1, FS2, FS1P, FS2P). Therefore, either of the frame sets provided enough coverage and diversity to span the relevant antimalarial chemical space.

## 6.2 Maps' performance: ActivityDB

All maps are exposed to the validation of their competence to predict the compounds that are active against the *Plasmodium* parasite. The validation is not completely unbiased, since the local map has been optimized for maximum performance in that exact task, but is nevertheless necessary in order to quantitatively evaluate the obtained maps.

First, let's consider the performance of the local map. While this challenge is not completely fair and biased towards the local map, since this is the end function of the manifold optimization procedure, this still provides a useful benchmark for this and all other maps. The performance scores (Figure 6.1) are good, and are comparable to previously reported SVM models [161] built on the same data — the mean BA for SVM models is 0.81, *versus* 0.80 for local GTM. However, the modeling paradigm is different in this case. SVM models were built individually, optimizing parameters and descriptor space separately for each data set. GTM classification, on the other hand, is based in this case on common manifold and common descriptor type for all data sets. The given manifold is 'colored' by a property using a training set, as described in the dedicated section, and the colored map serves to predict newly projected compounds of a test set. The cross-validation of a GTM model consists therefore in iteratively dividing available data into training and test set (here, in 3-fold cross-validation), and projecting them onto the given manifold to color and predict, until every molecule is predicted externally. Leaving

one-third of data out of coloring phase is quite challenging, but robust cross-validation results are obtained nevertheless. The goal of this simulation was not to maximize predictive power of each property-specific model, but to find a map supporting an optimal separation of actives *vs* inactives of all categories – and yet, results in terms of individual model performance are on par with dedicated SVM approaches. This is an excellent argument in favor of the robustness of the map.

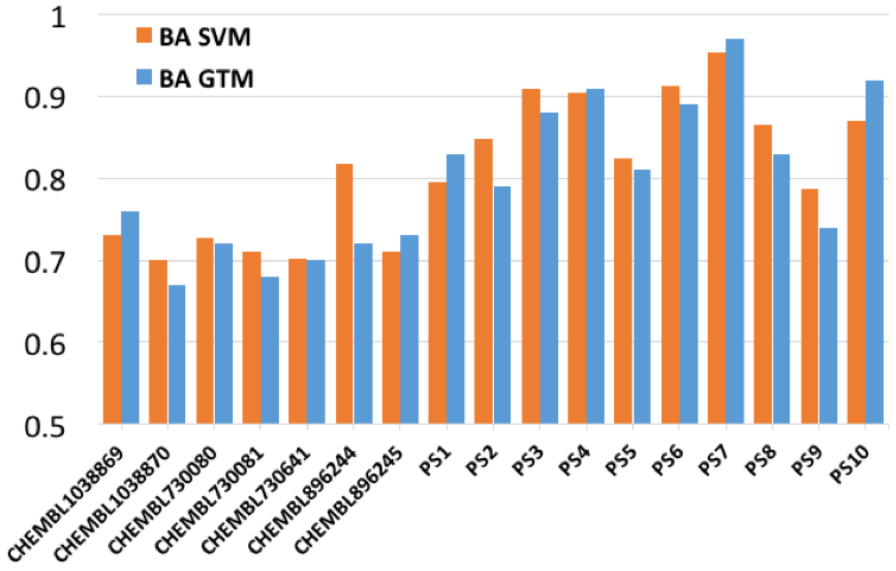



FIGURE 6.1: Performances of SVM (previously reported [161]) and GTM models in 3-fold cross-validated classification task with Balanced Accuracy (BA) score for 17 ActivityDB subsets. SVM models are built individually for each data set, GTM classification is based on common manifold, colored by 17 properties.

For the global maps of the antimalarial chemical space, this validation is more challenging. The end function of the map optimization in this case has been the effective separation of compounds by their respective target, not providing SAR information on specific series of compounds. The molecules of ActivityDB have been present during the manifold building procedure, since the whole MalariaDB is used as the frame set, and thus they do participate in the maps' chemical space definition. Nevertheless, the information on their antimalarial activity has never been introduced to the map, end therefore, this validation is external for these maps.

Figure 6.2 demonstrates the performances of the global and the local maps in external validation for 17 ActivityDB subsets. The global maps have, unsurprisingly, lower overall performance than the dedicated map: the average BA scores are 0.73, 0.72, 0.74, and 0.76 for Map1, Map2, Map3, and Map4, respectively. The Map4 has the closest score to the BA of the local map (0.80). As it can be seen from the figure, it has overall consistently good performance, and performs as well on subsets CHEMBL896244 and

CHEMBL896245, which have proved to be difficult for all other maps: Map3 even fails the challenge with the BA of 0.5 for both these subsets.

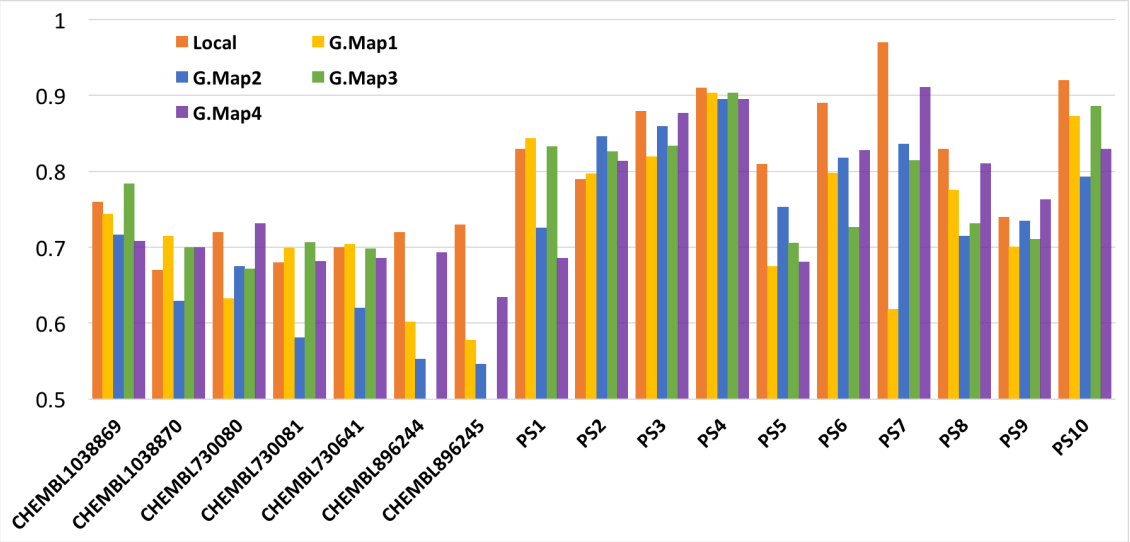



FIGURE 6.2: Performances of GTM global maps compared to the local map in 3-fold cross-validated classification task with Balanced Accuracy (BA) score for 17 ActivityDB subsets.

As for the universal map (Map2 has been chosen as it had the best overall score), this challenge is truly external: neither frame sets, nor color sets contain any information of antimalarial activity of a compound, whether it is the IC50 value or a mechanism of action. Figure 6.3 demonstrates the comparison of BAs of the universal and the local maps for 17 ActivityDB subsets. This map also struggles with CHEMBL730641, CHEMBL896244, and CHEMBL896245 subsets, but has overall good performance: average BA for the map is 0.74 *vs* 0.80 for the local map.

Hence, all the constructed maps preform well in terms of predicting active antimalarial compounds. Obviously, the dedicated local map has the highest scores that is comparable to classical SVM models. Nevertheless, even maps that were not optimized in active/inactive classification if antimalarials, but were still built in the chemical space relevant for the medicinal chemistry, are able to convey the knowledge to antiplasmodial activity.

## 6.3 Analysis of antimalarial chemical space

The analysis of the map from the point of view of chemical space can be done by the visual inspection of the given map. However, it is facilitated greatly by the application of the concept of responsibility patterns (RP) and privileged structural motifs (PSM). Most relevant privileged responsibility patterns, satisfying the hereby defined criterion

of CSP, are referring to a case where molecules of given pattern contribute to single node on the map, or, more rarely, to neighboring nodes, thus forming distinct clusters on the map. Such analysis may provide with suggestions of mode of action of antimalarials (as the neighborhood behavior principle suggests, molecules with identical RPs may correspond to the same mechanism of action or biological target), discover novel scaffolds that are worth to investigate further in more SAR studies, and so on. Here, we discuss the results obtained from the selected maps and the perspectives of the approach. PSMs are determined by visual inspection of compound sets matching each privileged RP.

### 6.3.1 ActivityDB on the local map

All available data of ActivityDB is used for PRP analysis of the local map. Since all the subsets concern experimental studies of antimalarial activity, and in all of them a fraction of compounds has been shown to kill the parasite (at different stages and under various conditions, thus the mechanism may not be the same), one may assume that if a molecule has been measured with a high antiparasitic activity in either of the 17 considered subsets, it is 'generically' considered active against the parasite. The objective of the exercise is the chemical space investigation: data sets may have overlaps in compound families, which should be studied from point of view of mechanism or species selectivity, while at the same time explore chemotypes tested up to this moment (as represented in ChEMBL database) and guide chemists to design new compound families, never used before against malaria.

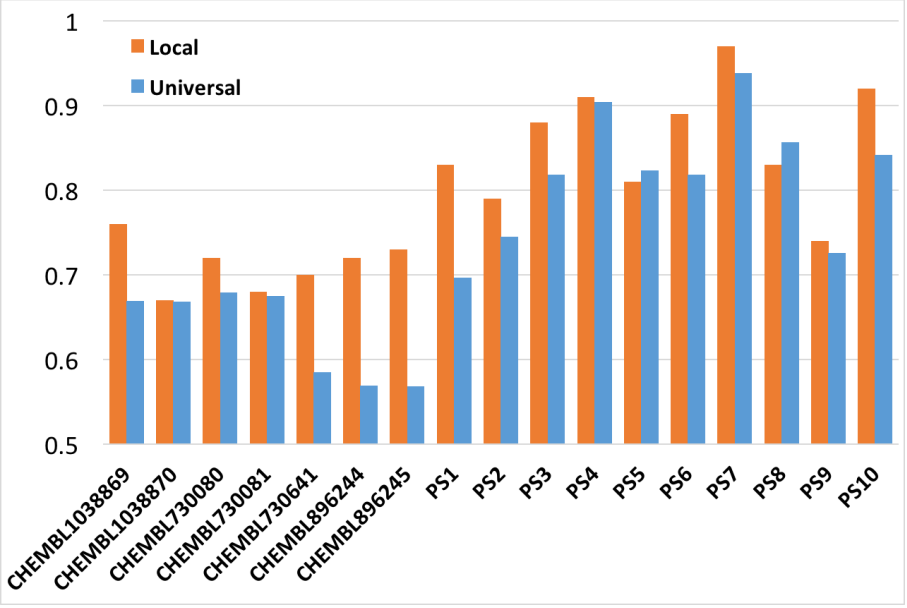



Figure 6.3: Performances of GTM universal map (Map2) compared to the local map in 3-fold cross-validated classification task with Balanced Accuracy (BA) score for 17 ActivityDB subsets.

The ACM representation of all available data is given on Figure 6.4. Several zones corresponding to most prominent PSMs (corresponding to high value of CSP > 20) are highlighted. Structural motifs themselves are shown in Table 6.2 (the numbers are following the zone numeration in Figure 6.4). As results of the analysis reveal, many data sets consider chloroquine analogues (4-aminoquinolines) for activity studies, which is not surprising, as chloroquine was one of the most efficient antimalarial drugs on the market. Smaller sets focus also on specific compounds series, such as artemisinin analogues: despite being the most used antimalarial nowadays, the family is present in only three data sets (PS5, PS7 and PS10), with PS7 being almost exclusive to this series. Another example is PS4, which contains a unique quinolone-based motif. Such set- and zone-specific patterns lead to distinction of mode-of-action-specific zones: since all artemisinin-like compounds are located in a small area, this area likely corresponds to the same mechanism as artemisinin itself.

Compound series associated to different bioactivity measures tend to privilege distinct RPs and hence different PSMs (see Table 6.2). One may therefore conclude about selective structural features required to provide activity under the various testing conditions, eventually supporting different mechanisms of action: several nodes populated by inactives in one data set are dominated by actives in another (e.g., PS2 and PS9 study activity against *Plasmodium K1* strain, while PS1 – against *Plasmodium* 3D7, so interspecies specificity of compounds may be suggested). Such interpretation is, however, unfortunately not fully supported by the data, since the entire activity matrix of test results of all compounds under all testing conditions is not provided. The default assumption that an

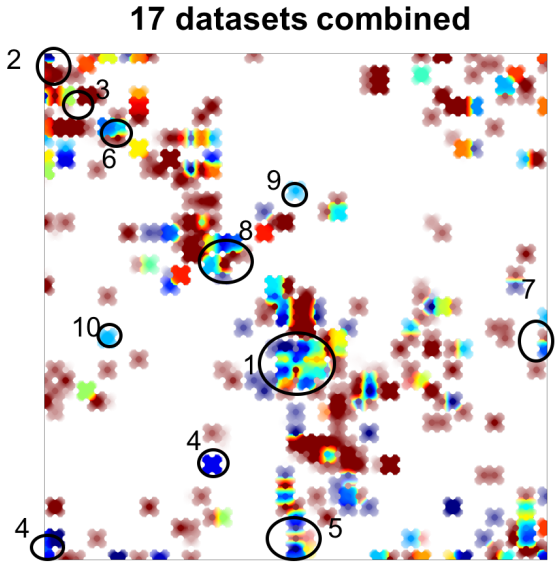



Figure 6.4: Local map with class landscape corresponding to ActivityDB annotations: blue for actives, red for inactives, intermediate color represent mixed population of the zone. Zone numeration is explained in Table 6.2

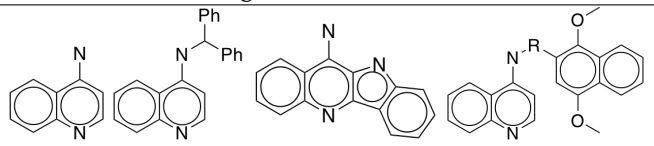
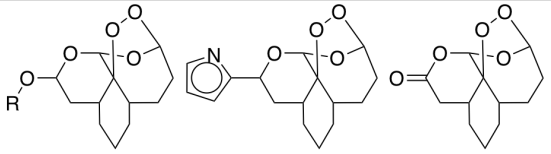
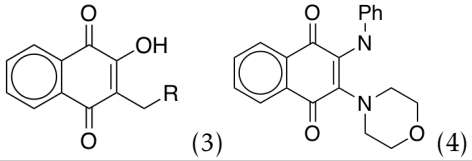
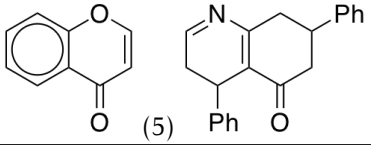
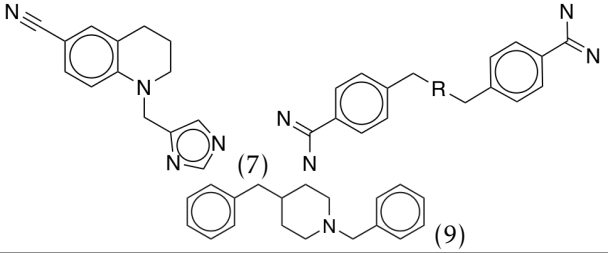
| Compounds | Privileged Structural Motifs | Native subsets |
|---|---|---|
| 4-amino-quinolines (1) |  | PS1, PS2, PS4, PS5, PS7, PS8, PS9 |
| Artemisinins (2) |  | PS5, PS7, PS10 |
| Naphtho-quinones |  (3) (4) | PS10, CHEMBL1038869, CHEMBL1038870, CHEMBL730080, CHEMBL730081 |
| Heterocyclic carbonyls |  (5) (6) | PS3, PS9, CHEMBL1038869, CHEMBL1038870, CHEMBL730081, CHEMBL730641 |
| Others |  (7) (8) (9) | PS9, CHEMBL896244 |

Table 6.2: Top privileged antimalarial structural motifs resulted from the analysis of responsibility patterns on the local map (the totality of ActivityDB). Numbers in parentheses correspond to the number of the zone on Figure 6.4.

active found under specific testing conditions would show up as inactive with a distinct testing protocol cannot be taken for granted. A 'privileged' pattern status simply means that the pattern is, statistically speaking, much more densely represented in a peculiar subclass of a library, with respect to the rest of the library. It does not imply a causal relationship from the associated PSM to the subclass membership (being active under a given test protocol), but may simply reflect some selection bias (compounds never picked for testing under a given protocol cannot possibly appear amongst the active subclass). Privileged pattern analysis – here and in the entire field of medicinal chemistry – is a tool meant to highlight imbalances of pattern densities in specific subclasses, but the underlying reason for such imbalances must be further investigated.

Certain PSMs are more prominent in bigger data sets, studying cytotoxicity of large, diverse sets: for example, naphthoquinone, flavone and quinolone analogues are privileged here, while are almost never represented in family-specific studies. Often, their mechanism is not clear, since they are measured in *in vitro* parasite proliferation and

large biochemical target screening assays, and often selected as false positives in high-throughput screens. Their unspecific interactions with numerous biological targets have resulted in the dogmatic view of PAINS (pan-assay interference non-specific compounds) [199]. Besides the example of atovaquone, a licensed antimalarial napthtoquinone derivative, which specifically targets the mitochondrial electron chain transfer at the $bc_1$ complex of *P. falciparum* parasites, certain menadione representatives are extensively studied in literature [200–202]. Many other 1,4-naphthoquinones are represented mostly in big, non-target and non-family specific data sets, with a MoA still under debate, and they are not present in smaller, scaffold-centric sets other than PS10.

Eventually, such a general representation of the data sets on the map provides insight into the chemical space of families or series of studied antimalarial compounds. As one notices, most of the nodes populated by active compounds are clustered together or are surrounded by nodes of the inactive class. Such clusters represent successful hit-to-lead optimization stories: a novel series of compounds is found to be active against malaria, and the chemical space is extensively investigated around that family. As an example, the center of the map (zone 1 in Figure 6.4) is populated by derivatives of 4-aminoquinoline, widely studied in antimalarial tests, and represent what may be called 'explored territory' in the chemical space of antimalarial compounds.

However, some active nodes on the map are isolated from others. These nodes are populated by small, chemically distinct series of compounds found in large screening collections, and constitute potential starting point for further investigation. For example, nodes 9 and 10 on the map (Figure 6.4) contain exclusively compounds from St. Jude screening set [162] (CHEMBL730080). Such isolated 'islands' constitute *terra incognita* of the available chemical space: first discoveries in these zones of chemical space are made, a hit is found, but the environment is yet to be explored, and leaves many possibilities for lead optimization around that island.

### 6.3.2 Global maps and antimalarial modes of action

The global maps, on the other hand, were trained specifically to separate the compounds of different modes of action. So, the analysis of chemical space of these maps is bound to reveal the meaningful structural information linked to the mechanisms and targets of antimalarial compounds.

The first layer of relevant information lies in the distribution of molecules of various mechanisms of action on the map. To further extend the available knowledge on the compounds, an additional class is considered for all of them – the target confidence level. To follow the previous disposition, two labels are assigned to each molecule –

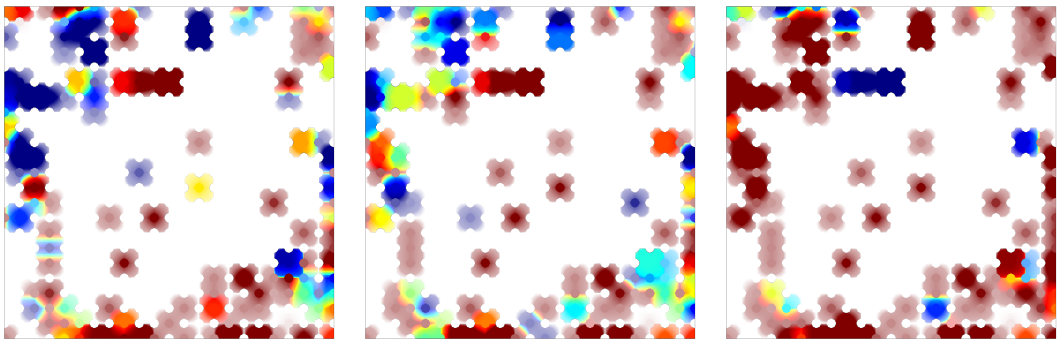


FIGURE 6.5: Global map G.Map1 with the class landscapes corresponding to kinase inhibitors (on the left) and compounds targeting the electron flow (on the right), blue color corresponds to compounds with this mode of action. In the center, the distribution of molecules by confidence level (blue for experimental, red for hypothetical).

Experimental or Hypothetical. This should help locate the zones where only compounds with non-confirmed target annotations are residing and guide chemists to these 'islands' that need yet to be explored experimentally.

On the Figure 6.5 global map G.Map1 is shown with three different class landscapes – inhibitors of kinases, inhibitors of electron flow in the parasite, and the distribution of compounds by confidence level. As expected, the inhibitors of kinases that are very structurally diverse are distributed more or less homogeneously all over the map. The inhibitors of the electron flow, on the other hand, form several distinct areas, since most molecules annotated by this precise group of modes of action belong to several specific chemical families.

The overlap of areas of certain MoA with the confidence level and the inspection of these zones may give another insight in the data distribution. Again, kinases are one of the most proliferous targets in medicinal chemistry, so the zones corresponding to kinase inhibitors often correlate with the zones of compounds with experimental annotations. Out of the few 'islands' of parasite electron flow disruptors, on the other hand, only a couple correspond to experimental confidence level. For example, as Figure 6.6 shows, a small zones in the top part of the map correspond to triazolopyrimidine-like structures, the inhibitors of DHODH [65], and most of these come from MMV MalariaBox project [123], where their annotations are confirmed experimentally. On the other hand, a larger area closer to the center of the map contains all sorts of substituted quinolones and pyridones []. Due to the structural diversity, these compounds occupy a big portion of the map. This type of compounds is believed to interact with the cytochrome $bc_1$ complex of the parasite, however, most of these have no experimental evidence of their mechanism of action.

The class landscape of kinase binders also provides the insight in the structural motifs that are privileged for this mechanism of action. Figure 6.7 illustrates several examples

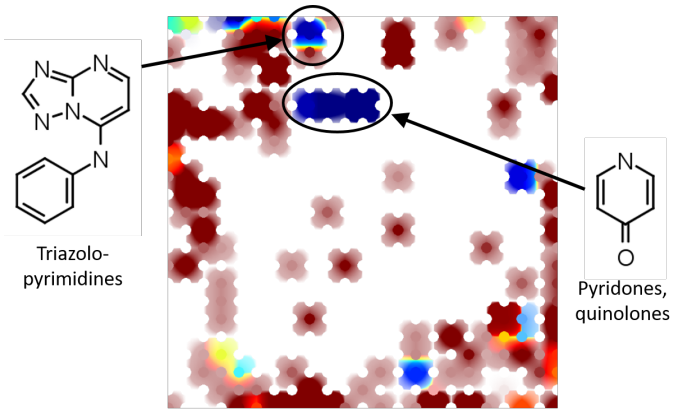



FIGURE 6.6: Global map (G.Map1) with the distribution of electron flow disruptors (blue for zones populated by molecules with this annotation, red for all others). Zones occupied by privileged structural motifs are highlighted.

of PSMs corresponding to the kinase inhibitors. Interestingly, all maps are able to recognize the underlying structural patterns despite the differences of the descriptor space: these (and others, not shown on the figure) privileged scaffolds are conserved across the four considered maps. However, the 4-aminoquinoline structure catches the eye here: it is generally believed that quinolines bind to heme rather than to specific protein. This and other examples of probably incorrect annotations are discussed in the following section.

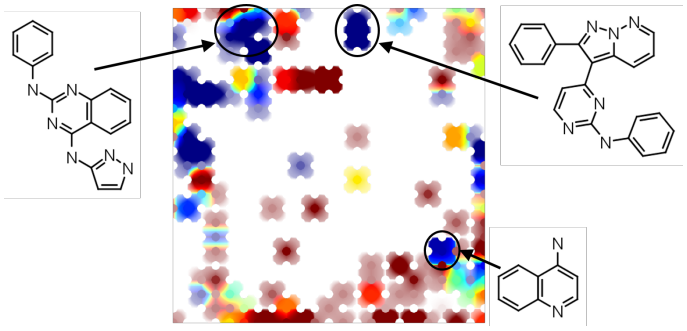


FIGURE 6.7: Global map (G.Map1) with the distribution of kinase inhibitors (blue for zones populated by molecules with this annotation, red for all others). Zones occupied by privileged structural motifs and corresponding PSMs are indicated.

### 6.3.3 Erroneous target annotations

Interestingly, that analysis of shared structural motives biased towards a certain mode of action sometimes leads to contradictory conclusions. There are cases when a particular motif is common for compounds that are annotated as ligands of different targets. Consequently, the question arises: which annotation should be taken as correct? The examples of such situation are discussed here.

As the first case, let us consider the kinase inhibitors. This is the most numerous category in MalariaDB, however, half of these annotations are hypothetical, which means that no experimental evidence of the mode of action of these compounds has been found. Figure 6.8 shows the G.Map1 with the distribution of kinase inhibitors (blue for zones populated by molecules that target kinases, red for all others). While mostly the kinase-populated zones correspond to compounds that share the kinase-specific scaffolds, one specific responsibility pattern corresponds to 4-aminoquinoline derivatives. This holds true also for the G.Map2 (Figure 6.10). But generally, 4-aminoquinolines target hemozoin formation in the parasite, so are these annotations incorrect? In this particular case, unfortunately, we cannot tell, because all molecules residing in this point have hypothetical annotations, and have aromatic heterocycles in the 7th position of the quinoline moiety, while Cl in that position is crucial for hemozoin formation inhibition according to literature [203]. Experimental verification of their mechanism of action is needed.

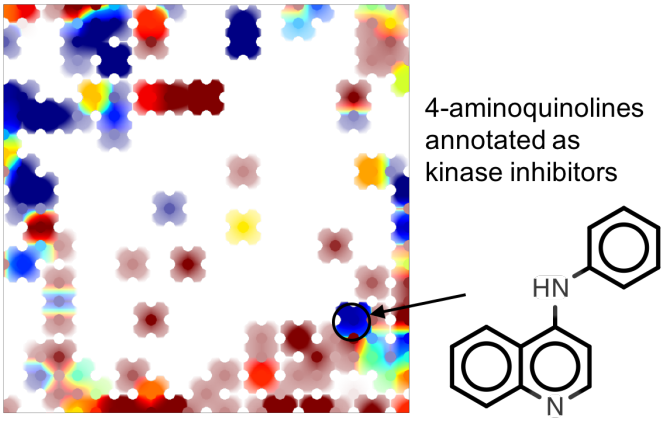



FIGURE 6.8: Global map (G.Map1) with the distribution of kinase inhibitors (blue for zones populated by molecules that target kinases, red for all others). Highlighted zones contains 4-aminoquinolines, generally believed to inhibit hemozoin formation, annotated as kinase inhibitors.

Other examples do allow us draw conclusions on the incorrectness of annotations. For instance, if we inspect the G.Map1 colored by Hemozoin digestion MoA (Figure 6.9), we can find a "mixed" zone on the right. While the color zone is very close to red (meaning that most compounds in there do not correspond to this specific mechanism), this is because most of the molecules in there have hypothetical MoA annotations, majoritarily as kinase inhibitors. However, several compounds in the area are experimentally confirmed to have hemozoin formation inhibition mechanism, so that may be an indication for MoA correction of molecules without experimental annotation.

The same situation arises in Map2 (colored by kinase MoA, Figure 6.10). The indicated motif corresponds to two different annotations: experimentally confirmed 'Hemozoin formation inhibition' (in [162]), and hypothetical 'Kinase' (by Gamo et al. [119]). Again, it may be an indication to modify and suggest the MoA for all those molecules. Of

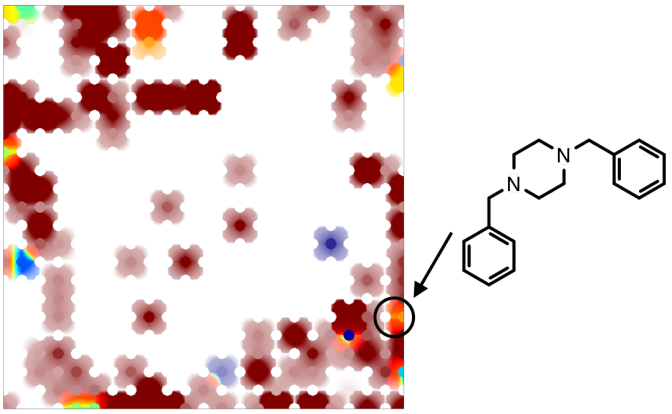



FIGURE 6.9: Global map (G.Map1) with the distribution of hemozoin digestion inhibitors (blue for compounds annotated by this mechanism, red for other mechanisms). A zone with mixed annotations and corresponding structural motif are indicated.

course, experimental validation is needed to confirm or deny the incorrect annotation hypothesis.

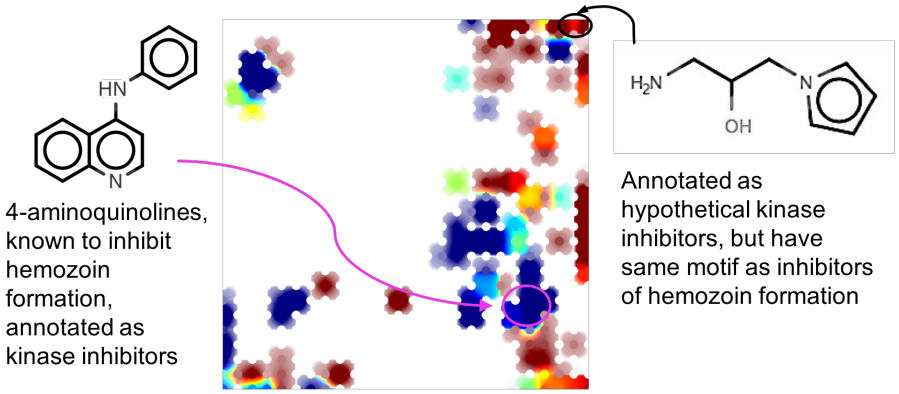



FIGURE 6.10: Global map (G.Map2) with the distribution of kinase inhibitors (blue for compounds annotated as kinase inhibitors, red for others). Two zones corresponding to possible incorrect annotations and corresponding structural motifs are highlighted.

## 6.4 Conclusion

The high polyvalence of Generative Topographic Mapping, which is both, a tool for visualization, clustering and rationalization of chemical information, and a tool for predictive modeling allows to conduct a deep analysis of the chemical space of any compound library. In the difficult field of antimalarial activity, where mechanistic information is sparse, while primary testing results stem from various, not always directly comparable testing protocols, GTMs provide a unique opportunity to bring together all this partial, complementary data into a common conceptual framework. It has been shown that such a common conceptual framework, represented by an ensemble of maps of varying degree of generalization, does exist. Several levels of maps of antimalarial chemical

space have been constructed: local maps for maximum predictive power, with respect to all the various antimalarial activity measures; global maps, effectively separating the compounds of different modes of action; and universal map, encompassing the general drug-like chemical space.

17 GTM-based classification models were derived from the respective key sets of ActivityDB corresponding to different anti-*Plasmodium* activity assessment protocols. They perform similarly to previously reported SVM models in both cross-validation and external test set prediction for the three kind of maps.

Two dimensional GTM maps well separate active and inactive compounds in distinct clusters populated by molecules with particular chemotypes. The data distribution is visualized in form of class landscapes providing with an overview of data coverage of the map with density-dependent color intensity modulation. Such a representation is intuitive for medicinal chemists: zones populated by active or inactive compounds are easily identified and may guide chemist in the search of the important structural features in the lead optimization process. Additionally, isolated 'islands' populated by actives represent structurally novel hits found in large screens, which would be interesting starting points for more detailed SAR studies.

Eventually, the maps can be used to investigate the mechanism-specific chemical space, by the application of privileged responsibility patterns approach. This approach allows to distinguish zones and corresponding structural motifs responsible for a given mode of action, as well as to detect the inconsistencies of data annotation. Several case of such erroneous annotation have been pinpointed, and their further investigation is recommended.

J Comput Aided Mol Des
DOI 10.1007/s10822-017-0019-4

# QSAR modeling and chemical space analysis of antimalarial compounds

**Pavel Sidorov[1,2] · Birgit Viira[3] · Elisabeth Davioud-Charvet[4] · Uko Maran[3] · Gilles Marcou[1] · Dragos Horvath[1] · Alexandre Varnek[1]**

Received: 21 December 2016 / Accepted: 18 March 2017


**Abstract** Generative topographic mapping (GTM) has been used to visualize and analyze the chemical space of antimalarial compounds as well as to build predictive models linking structure of molecules with their antimalarial activity. For this, a database, including ~3000 molecules tested in one or several of 17 anti-*Plasmodium* activity assessment protocols, has been compiled by assembling experimental data from in-house and ChEMBL databases. GTM classification models built on subsets corresponding to individual bioassays perform similarly to the earlier reported SVM models. Zones preferentially populated by active and inactive molecules, respectively, clearly emerge in the class landscapes supported by the GTM model. Their analysis resulted in identification of privileged structural motifs of potential antimalarial compounds. Projection of marketed antimalarial drugs on this map allowed us to delineate several areas in the chemical space corresponding to different mechanisms of antimalarial activity. This helped us to make a suggestion about the mode of action of the molecules populating these zones.

**Keywords** Antimalarial compounds · Quantitative structure–activity relationships (QSAR) · Generative topographic mapping (GTM) · Chemical space · Mode of action

**Abbreviations**

| | |
|---|---|
| (Q)SPR/SAR | (Quantitative) structure–property/structure–activity relationships |
| (P)RP | Privileged responsibility patterns |
| PSM | Privileged structural motifs |
| GTM | Generative topographic map |
| AD | Applicability domain |
| MoA | Mode of action |

## Introduction

Malaria remains one of the most severe infectious diseases that disproportionally affects the public health and economic welfare of the world's poorest communities. The causative agents of malaria belong to five species of protozoan parasites of the genus *Plasmodium* [1]. The most prominent and dangerous killer among them, *Plasmodium falciparum*, is responsible for most severe forms of the disease, such as cerebral malaria. Also, *P. vivax* malaria represents an important public health challenge because it has a wider geographical distribution than *P. falciparum* and the parasite can develop a dormant liver stage (known as a hypnozoite), causing a relapse of symptoms upon activation,

**Electronic supplementary material** The online version of this article (doi:10.1007/s10822-017-0019-4) contains supplementary material, which is available to authorized users.

✉ Dragos Horvath
dhorvath@unistra.fr

✉ Alexandre Varnek
varnek@unistra.fr

[1] Laboratoire de Chemoinformatique, UMR7140 CNRS-Université de Strasbourg, 1 rue Blaise Pascal, 67000 Strasbourg, France

[2] Laboratory of Chemoinfomatics, Butlerov Institute of Chemistry, Kazan Federal University, Kazan, Russia 420008

[3] Institute of Chemistry, University of Tartu, 50411 Tartu, Estonia

[4] Bioorganic and Medicinal Chemistry Team, European School of Chemistry, Polymers and Materials (ECPM), UMR 7509 CNRS—University of Strasbourg, 25, rue Becquerel, 67087 Strasbourg, France

Published online: 03 April 2017

months after an initial infection. Most malaria cases in 2015 are estimated to have occurred in the WHO African Region (88%), followed by the WHO South-East Asia Region (10%) and the WHO Eastern Mediterranean Region (2%) [2]. Previously limited to tropical countries, in particular to Sub-Saharan Africa, the disease is now progressing in non-endemic regions [3] due to both global warming and migration flows. According to the WHO, 214 million clinical cases of malaria have been reported worldwide in 2015 giving rise, to 438,000 deaths per year that represents a 48% decline in mortality since 2000. In 2015, whereas malaria is still the fourth highest cause of death, accounting for 10% of child deaths in sub-Saharan Africa, an equally impressive drop in malaria-related morbidity has been observed [2]. These spectacular statistics are partly due to the massive use of insecticide-treated bednets (ITNs) and artemisinin-based combination therapies (ACTs). In sub-Saharan Africa, of the million cases averted due to malaria control interventions it is estimated that 69% were accounted for the use of ITNs, 21% for ACT and 10% for indoor residual spraying. Unfortunately, decreases in case incidence and mortality rates were slowest in countries that had the largest numbers of malaria cases and deaths in 2000. Reductions in incidence need to be greatly accelerated in these countries if global progress for malaria elimination is to improve [2]. Furthermore, since 2008 there has been worrying decrease in the efficacy of ACT treatment in South-East Asia due to the emergence of drug-resistant parasites, which endangers the recent achievements and threatens the world's malaria control and elimination efforts. In 2015 the Global technical strategy for malaria 2016–2030 was adopted by the World Health Assembly to target reductions in malaria cases and deaths for approaching malaria eradication. To reach this ambitious challenge in areas affected by ACT and multidrug resistance, the strategy prioritizes the rapid interruption of transmission of parasites to mosquitoes by new drugs targeting essential metabolic pathways in the insect stages. This is of critical importance, given the increased fitness of ACT resistant parasites to persist for days to weeks in a non-replicating state and to survive drug exposure long enough before recovery and transmission [4]. The licensed 8-aminoquinoline drug primaquine [5] and the blue dye—methylene blue [6]—known as the first synthetic antimalarial molecule discovered in the early twentieth century, possess transmission-blocking properties, but these drugs, in addition to other liabilities, produce either hemolytic anemia in individuals with glucose-6-phosphate deficiency or unsatisfying in vivo efficacy when used alone in humans, respectively.

Since the complete genome sequencing of *P. falciparum* in 2002, partial genome sequencing of other *Plasmodium* species (*P. yoelii, P. vivax* and P. knowlesi) and various species races of Anopheles gambiae, the mosquito that transmits *P. falciparum* infection, malaria research has reached the post-genomic era [7]. Identification and elucidation of drug targets from the genome might be retrieved to enable an acceleration of preclinical candidates into the drug and vaccine pipelines. Genome sequences allowed increased knowledge of the basic biology of malaria parasites, and opened new avenues of research in chemobiology to lead to new, rationale, therapeutic approaches. Chloroquine, the popular, licensed and widely used antimalarial drug in the first half of twentieth century, had known restricted use because of the significant loss of efficacy in resistant parasite strains worldwide [8]. A plethora of 4-aminoquinoline analogues and derivatives, and many other classes of compounds have been developed since, and studied to overcome the drug resistance. Although many existing antimalarial drugs currently work well against malaria, the precise biological targets that they affect are often not entirely known. Novel genetic approaches [9] associated to phenotypic and chemoinformatic [10] screening can be used to discover new targets quickly and successfully, with the ultimate goal of finding new antimalarial chemical entities.

In silico methods help to accelerate and facilitate the discovery and design of novel active molecules. However, such investigations demand sufficient amounts of high-quality experimental data, and in order to be exhaustive it should include as many chemotypes as possible. Currently, most studies available in the literature report quantitative structure–activity relationships (QSAR) for antimalarial compounds of one family at a time in a consistent biological experiment: substituted 4-aminoquinolines [11–13], naphthoquinone derivatives [14], analogues of other used drugs [15–17], using different methods and molecular descriptors. Gamo et al. [18] widened the investigation by screening a large collection (millions) of compounds to identify active molecules, while proposing an overview of the chemical space of antimalarial compounds (i.e. the span of active chemotypes), as well as proposed mode of action hypotheses for some of these families. However, this overview was limited to the studied collection and did not include several critical antimalarial drugs (such as artemisinin) for validation and verification. Nevertheless, this assay data has been used in several works for both clustering and prediction of biologically relevant antimalarial targets and SAR information (by genetic methods [19] and network graph analysis [20, 21]). Most recently, an open-source project for antimalarial drug discovery and investigation—MalariaBox [22]—has been initiated. The project assembles 400 diverse drug-like molecules for which antimalarial activity and mechanism of action are studied experimentally.

The current study is dedicated to an analysis of structure-antimalarial activity data available both publicly, in

the ChEMBL database [23] and in-house (Dr. Davioud-Charvet's group). ChEMBL has a dedicated Malaria Data subset and is therewith one of the main publicly available structure–activity data repositories dedicated to malaria. Reported data is unfortunately quite heterogeneous and fragmented into many subsets of widely varying sizes (from <100 to $10^4$ entries/subset) tested against different parasite strains at different stages of parasite development and with significantly different screening conditions. The in-house library (Dr. Davioud-Charvet's group) contains 266 1,4-naphthoquinones with known chemical, physico-/electro-/bio-chemical and biological properties [24–26]. An intensive data curation procedure is applied, in order to regroup data from comparable testing protocols into common sets, by ignoring noisy data or screening campaigns too poor in active compounds. For analysis purposes, Generative topographic mapping, previously successfully used in various QSAR challenges [27], is utilized in order to build predictive models, as well as to investigate the chemical space of data of high confidence available in ChEMBL for molecules active against *P. falciparum* parasite. The resulting models, with performance comparable with other machine learning methods, not only provide robust predictions of compound activity in different experimental conditions, but also outline densely populated and empty zones of the compound space, indicating which chemotypes are not yet extensively studied for their activity. Some suggestions about the putative mode of action may also be drawn from the map.

## Data and methods

### Data

Two primary sources of structures and experimental values of antimalarial activity (i.e. activity against the human pathogen *P. falciparum*) have been used in the current work. First, the experimental data from the Laboratory of Bioorganic and Medicinal Chemistry (University of Strasbourg): 266 antimalarial compounds were recently synthesized and antimalarial activity was measured, using ten testing protocols, under varying conditions. Second, several with reported in vivo antimalarial activity have been extracted from the ChEMBL database (https://www.ebi.ac.uk/chembl, version 20).

The main approach was to complete in-house data with publicly reported bioactivity values obtained by the same or similar protocols. Only compounds with reported $IC_{50}$ values were considered. As a result, 2295 compounds were added to antimalarial database. The second approach was searching ChEMBL database by target, retaining compounds for which different activity types ($IC_{50}$, $EC_{50}$, inhibition %, etc.) were reported. *Plasmodium* was assigned as target, which resulted in extraction of 249,658 (250 K) compounds and 400,176 values for 2900 functional assays. In order to reduce the number of modeled properties and compounds, different protocol curation and merging strategies were used, as described elsewhere [28]. As a result, 30 compound series associated to distinct testing protocols were obtained. Seventeen 'key' data sets containing 2955 molecules in total, are shown to be modelable (leading to statistically robust SVN classification models), and consequently have been selected for further investigation in this paper. Class labels are based on the reported activity for all molecules: if compound has its activity value (EC50, IC50, etc.) in submicromolar range, it is annotated as 'active', and 'inactive' otherwise. Detailed information on experimental conditions, studied biological property and composition of these datasets is available as Supplementary information (see Supplementary materials, Table 1s).

In addition, 36 known antimalarial drugs were hand-picked to constitute an external validation set, as well as to form a set of mechanism-annotated molecules to delineate "mode of action"-specific areas on the map.

#### *Frame sets*

The evolutionary algorithm (see "Generative topographic mapping" section) was enabled to select the optimal choice out of four different frame sets considered as working hypotheses for manifold fitting. Two of these, labeled FS1 and FS2, are (disjoined) random picks from the global set of 250 K Malaria-associated ChEMBL compounds, of 10 K compounds each. No further checking of the activity flag or conditions of measurement of their activity was undertaken. They are thought to represent the chemical space considered by experts to be relevant for antimalarial research—i.e. compounds considered by experts to be worth testing against *Plasmodium*. Alternatively, two additional working hypotheses labeled FS1P and FS2P respectively were obtained by adding, to FS1 and FS2 respectively, 1700 random activity-annotated compounds from ChEMBL series with dose-response data ($pIC_{50}$, $pEC_{50}$), in order to ensure that large high-throughput screening libraries tested at a single concentration do not completely control the frame sets.

#### *Molecular descriptors*

Compound standardization followed the default protocol [29] installed on our public web server, and powered by ChemAxon [30, 31] tools. It includes removal of very large entities (>100 heavy atoms), counter-ion strip-off, split-charge representation of N-oxides, basic aromatization,

conversion to the most populated microspecies of the most probable tautomeric form, etc.

A pool of 39 diverse ISIDA fragmentation schemes [32–35] served as initial choices for the best suited descriptor spaces for the map of antimalarial compounds. They include sequences, atom pairs, circular fragments and triplet counts, colored by atom symbols, pharmacophore features or force field types. These 39 fragmentation schemes were selected for their relatively low number of fragments they generate (less than 3000 each).

## Generative topographic mapping

Generative topographic mapping (GTM) [36] is a method of non-linear mapping that has been successfully used in various domains of data analysis.

In GTM, each point in the low-dimensional (usually 2D) latent space is mapped onto the manifold embedded in the initial descriptor space. The manifold has a topology of a square grid of $K \times K$ nodes, and the mapping function $y(x, W)$ is defined as a mixture of $M \times M$ radial basis functions (RBFs) of width $w$ forming a regular grid in the latent space. For each data point (molecule) $n$, its probability to be assigned (associated) to a manifold grid node $k$ is called responsibility $R_{kn}$, and is normalized (the sum of probabilities of having the molecule reside somewhere on the map is $\sum_{k=1}^{K} R_{kn} = 1$). Vectors of responsibilities of molecules are then used for visualization purposes, as well as for building GTM-based classification and regression models.

### *GTM classification models and applicability domain*

GTM classification models [37] are relying on the neighborhood principle. Training set molecules are placed onto the map, and nodes are colored by dominating class (i.e. molecules of which class contribute most into the given node). Next, the molecules of the test set are projected to the colored map, and their class is then predicted by kNN or a Bayesian approach based on their responsibility distribution. The visualization and analysis of obtained maps is done through Activity Class Map (ACM). The ACM is an image of the square grid of GTM, where color intensity of a node reflects the data density of that node. Color choice reflects the dominant class of the node—a node is said *active* if the sum of active compound responsibilities exceeds the one of inactives. GTM-based regression models are not in the scope of this paper and are discussed elsewhere [38].

GTM applicability domain (AD) may be defined in several manners [37–39]. In this work, only data density-based applicability domain is considered. When the ACM is built, a certain threshold can be applied for data density: if molecules of the training set have near-null responsibility in a node, that node will not contribute in class or value prediction, thus it may be considered as out of AD.

### *GTM parameter optimization*

The result of the mapping will depend on a number of parameters, both internal for the method (such as the number and width of RBFs and the topology of the manifold defined by the number of nodes), and external, for example, the initial data set and descriptor space. The obvious and straightforward approach to optimization [39] of the map performance is to scan the possibilities evaluating the model performance in cross- or external validation. However, the enumeration of all possible maps would cause a combinatorial explosion, especially since some of the parameters take continuous values. Therefore, map parameters cannot be fitted by systematic search. Their optimization is based on evolutionary optimization procedure [40] and takes advantage of GTM model learning paradigm: since manifold building and projection and prediction steps are distinct, different data sets may be used in these steps.

Input provided to the algorithm consists of one of considered frame sets, encoded in one of the considered descriptor spaces, and a set of method parameters, according to the current 'chromosome' (vector of the degrees of freedom) of the evolutionary procedure. The manifold is built in an unsupervised manner (i.e. there is no need to assign a property to compounds of the frame set) to outline the chemical space of the map. Next, selection sets with annotated properties are projected onto the frame and are exposed to cross-validated modeling process. In this work, only classification models were considered. For each selection set, a model performance measure is calculated. Here it is the balanced accuracy:

$$BA = 0.5\left(\frac{TP}{TP + FN} + \frac{TN}{TN + FP}\right)$$

where TP is the number of correctly predicted actives, TN—the number of correctly predicted inactives, and FP and FN are the number of incorrectly predicted inactive and active compounds, respectively.

Finally, the map (i.e. the combination of frame set, descriptor space and set of GTM parameters) is scored by the mean performance of all selection set-specific models, penalized by the standard deviation of individual model performance (BA). This 'fitness' score guides the evolutionary process in the space of 'chromosomes'—which represent de facto recipes for GTM construction. Eventually, top-scoring maps are exposed to external validation procedure. Schematically the optimization procedure is shown on Fig. 1.

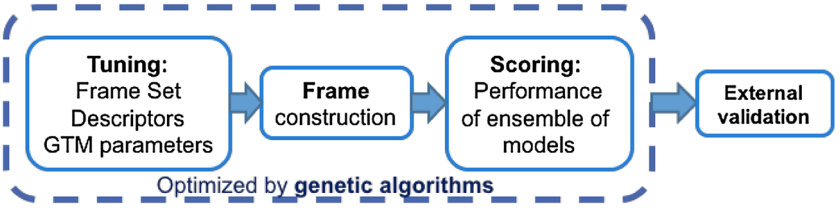



**Fig. 1** Schematic representation of the optimization algorithm used in the current work. From tunable parameters (frame set, descriptors, method parameters) a frame is built unsupervisedly. Next it is scored by the average performance of an ensemble of models. When an optimal map is obtained, it is exposed to external validation

*Privileged pattern detection*

As implied by the neighborhood principle, two molecules with identical or similar responsibility vectors (i.e., located closely on the map) should have similar properties. However, responsibility vectors consist of real values, thus detecting molecules with strictly identical responsibilities is highly unlikely. Therefore, it has been proposed [41] to transform the vector into a discretized form, where responsibilities would be binned by intervals of 0.1, and assigned levels 0–10. Such discrete representation is called "responsibility pattern" (RP) of a molecule. Molecules with the same RPs are considered to belong to a same cluster or 'family', because they share some common structural motif which is the underlying reason for their mapping into the same map zone 'covered' by the RP. As already shown [41], common structural motifs may range from precisely defined scaffolds or even specifically substituted scaffolds, to fuzzier ensembles of related, interchangeable scaffolds, to even fuzzier 'pharmacophore-like' patterns.

If an RP cluster appears to be enriched in active compounds, then the common structural motif defining the cluster is privileged with respect to that property. Quantitatively that can be measured as relative frequency of specific RP for given activity—class specificity (CSP):

$$CSP(RP,\ C) = \frac{f_{actives\ in\ class\ C}(RP)}{f(RP)}$$

where $f_{actives\ in\ class\ C}$ (RP) is the occurrence frequency of RP among 'active' compounds (*vide infra*), and *f*(RP) is the pattern occurrence frequency within the set of compounds used to build the manifold (i.e. frame set). In this work, class would correspond to activity annotation in a combined set: a molecule is considered 'active' against the parasite if it is measured as active in any of 17 key sets, and inactive otherwise. This analysis only makes sense for often encountered responsibility patterns: in practice, only RPs represented at least ten times in the data set were considered for analysis. Such analysis may provide with suggestions of mode of action of antimalarials (molecules with identical RPs may correspond to the same mechanism of action or biological target). The most prominent privileged RPs are discussed if CSP reached 20 (i.e. they occur 20 times more often among actives of the combined set than in the frame set of compounds).

## Results and discussion

### Optimal antimalarial map

As a result of the optimization procedure, the map with the best overall performance, i.e. capable of simultaneously separating active and inactive compounds in 17 property classification tasks, is selected for further analysis. In the map selection process, manifold construction per se is actually unsupervised learning (unaware of activities), and the selection criterion is based on aggressive cross-validation and on simultaneous consideration of several properties. It was previously shown [39] to produce maps with excellent external validation capacities, extending to compound sets and properties completely different to the ones used for selection. Due to the shortage of modelable antimalarial structure–activity data sets, no such external validation could be envisaged here. However, the three-fold cross-validation results are already a robust proof of statistical soundness.

The size of the map is 49×49 nodes, and the descriptor type selected by the genetic algorithm as the best initial space is atom-centered fragments of length 1–2, with information on both atoms and bonds, and colored by force-field types. Thus, fragments retain most information about relevant chemical environment of all atoms in a molecule. The map has shown good performance in threefold cross-validated classification task for all 17 key data sets, with average balanced accuracy score of 0.80. Distinction between active and inactive compounds is more difficult (resulting in lower BA) for diverse sets of in vivo cytotoxicity studies.

The optimal frame set was FS2P, featuring the additional 1.7 K compounds with dose-response activity data. Yet, only 42 of the latter are present within the 17 key compound sets. This negligible overlap is not an important

factor, since (a) the overlap is small and (b) maps of near-equal statistical robustness (results not shown) could be obtained on the basis of all four frame set choices (FS1, FS2, FS1P, FS2P). Therefore, either of the frame sets provided enough coverage and diversity to span the relevant antimalarial chemical space.

These performance scores are comparable to previously reported SVM models [28] built on the same data (Fig. 2). However, the modeling paradigm is different in this case. SVM models were built individually, optimizing parameters and descriptor space separately for each data set. GTM classification, on the other hand, is based in this case on common manifold and common descriptor type for all data sets. The given manifold may be 'colored' by a property using a training set, and the colored map serves to predict newly projected compounds of a test set. The cross-validation consists therefore in iteratively dividing available data into training and test set (in ratio 2–1), and projecting them onto the given manifold to color and predict, until every molecule is predicted externally. Leaving one-third of data out of coloring phase is quite challenging, but robust cross-validation results are obtained nevertheless. The goal of this simulation was not to maximize predictive power of each property-specific model, but to find a map supporting an optimal separation of actives versus inactives of all categories—and yet, results in terms of individual model performance are on par with dedicated SVM approaches. This is an excellent argument in favor of the robustness of the map.

Note that the same 17 key data sets were already used as external challenge sets on 'universal' maps spanning the entire drug-like space, and the therein achieved balanced accuracy of active versus inactive compounds separation was significant (Challenge #6 in [39]), but slightly

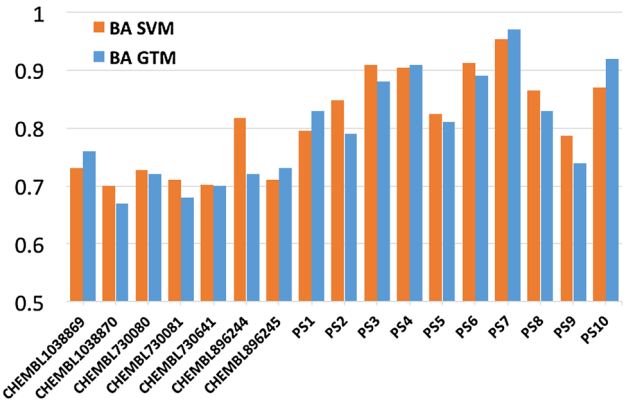



**Fig. 2** Performances of SVM (previously reported [28]) and GTM models in threefold cross-validated classification task with Balanced Accuracy (BA) score for 17 key data sets. SVM models are built individually for each data set, GTM classification is based on common manifold, colored by 17 properties

lower than in the present work. This is not surprising, as the frame set—even though containing only a minority of the key data set compounds—had the advantage to focus on malaria-relevant molecules only, rather than on samples of binders to radically different targets.

Since all the data sets concern experimental studies of antimalarial activity, and in all of them a fraction of compounds has been shown to kill the parasite (at different stages and under various conditions, thus the mechanism may not be the same), one may assume that if a molecule has been measured with a high antiparasitic activity in either of the 17 key sets, it is 'generically' considered active against the parasite. The definition of 'active' versus 'inactive' in each of the 17 key sets was the same used for previously reported classification models [28]. Therefore, a general two-class classification landscape ('active'—in at least one key set, versus 'inactive'—in all sets to which it belongs) is presented in Fig. 3, left side. In that case, the balanced accuracy is 0.74, thus slightly lower than average for individual sets, although it is not surprising, considering the diversity of studied chemotypes and the generalized character of prediction. The resulting map spans the studied chemical space of antimalarial compounds and summarizes the knowledge about structure–activity relationships learned from the totality of antimalarial data available in these sets. By contrast, some data set-specific ACMs are also illustrated in Fig. 3. Since density-based color intensity is applied, blank areas correspond to nodes where no significant population of any given data set members is observed. All key compound data combined constitutes roughly one-quarter of the used frame set in size, thus it would not cover the whole map (see Supplementary material, Fig. 1s). It is interesting to also note the difference between chemical space coverage of sets (as illustrated by number and position of colored nodes on the map) depending on their size and diversity. Smaller sets, mostly investigating limited series of compounds. For example, the only actives found in PS1 are 4-aminoquinolines. Such sets form small clusters on the map, separating effectively active and inactive molecules of different chemotypes present in the set. Big diverse sets, mostly related to cytotoxicity studies (e.g. CHEMBL730080), cover large portion of the map. Nevertheless, the performance of models concerning these bigger sets is satisfactory, since GTM prediction is based on responsibility distribution of predicted molecules over the map, and compounds are often dispatched over small cluster of neighboring nodes.

## Privileged pattern analysis

The overlap and selectivity of chosen data sets can also be studied from the point of view of privileged responsibility patterns. Most relevant privileged responsibility patterns,

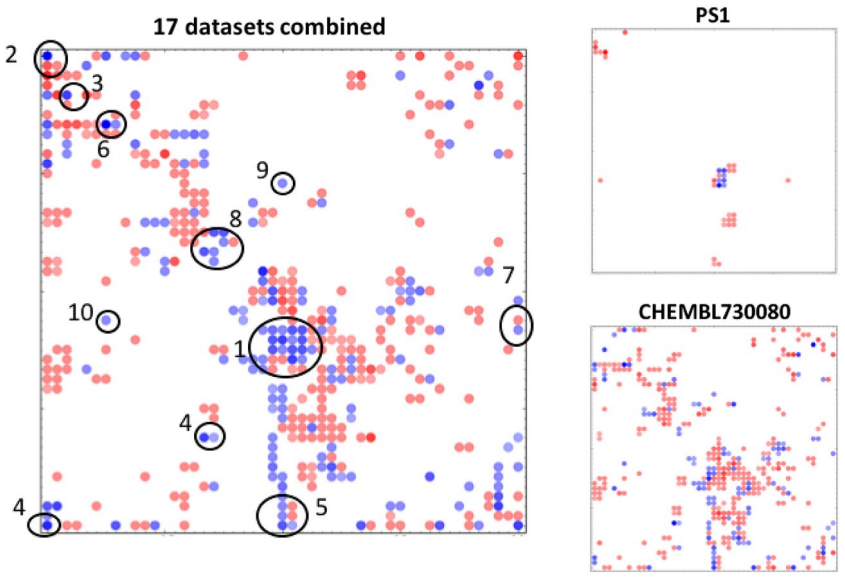



**Fig. 3** ACM representations of all available data (*left*), PS1 (*top right*), and CHEMBL730080 (*bottom right*) data sets. *Blue nodes* are mostly populated by active compounds, *red one* by inactives. *Smaller sets* limited to one series of compounds form distinct clusters of active and inactive compounds, while *big diverse sets* cover almost all chemical space. Zone numeration (*left*) is explained in Table 1

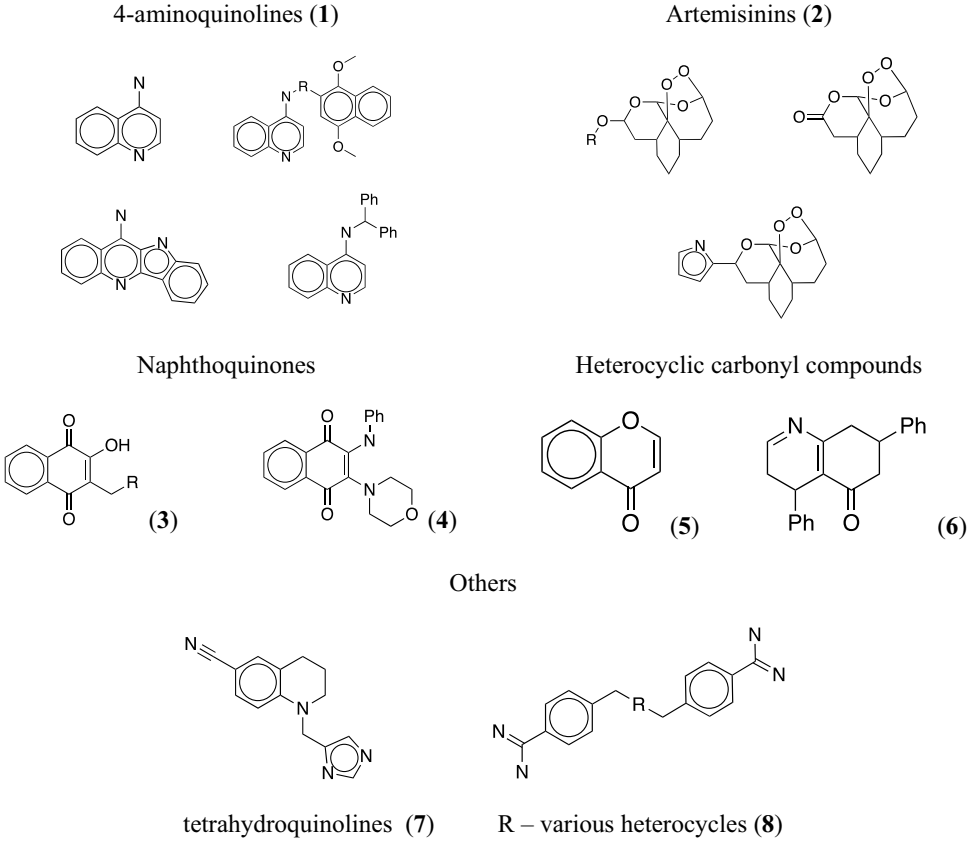
satisfying the criterion of CSP > 20, are referring to a case where molecules of given pattern contribute to single node on the map, or, more rarely, to neighboring nodes, thus forming distinct clusters on the map. It is worth to note that such clusters correspond often to molecules that share common privileged structural motifs, (PSMs—which may, but do not have to, be common *scaffolds*). Therefore, the analysis allows investigating the chemical space coverage of different data sets as well. PSMs were determined by visual inspection of compound sets matching each privileged RP.

Compound series associated to different testing protocols tend to privilege distinct RPs and hence different PSMs (see Table 1). One may therefore conclude about selective structural features required to provide activity under the various testing conditions, eventually supporting different mechanisms of action: several nodes populated by inactives in one data set are dominated by actives in another (e.g., PS2 and PS9 study activity against *Plasmodium* K1 strain, while PS1—against *Plasmodium* 3D7, so interspecies specificity of compounds may be suggested). Such interpretation is, however, unfortunately not fully supported by the data, since the entire activity matrix of test results of all compounds under all testing conditions is not provided. The default assumption that an active found under specific testing conditions would show up as inactive with a distinct testing protocol cannot be taken for granted. A 'privileged' pattern status simply means that the pattern is, statistically speaking, much more densely represented in a peculiar subclass of a library, with respect to the rest of the library. It does not imply a causal relationship from the associated PSM to the subclass membership (being active under a given test protocol), but may simply reflect some selection bias (compounds never picked for testing under a given protocol cannot possibly appear amongst the active subclass). Privileged pattern analysis—here and in the entire field of medicinal chemistry—is a tool meant to highlight imbalances of pattern densities in specific subclasses, but the underlying reason for such imbalances must be further investigated.

Eventually, all available data should be used for PRP analysis. In this case, the objective of the exercise is the chemical space investigation: data sets may have overlaps in compound families, which should be studied from point of view of mechanism or species selectivity (see above), while at the same time explore chemotypes tested up to this moment (as represented in ChEMBL database) and guide chemists to design new compound families, never used before against malaria.

The ACM representation of all available data is given on Fig. 3, left side. Several zones corresponding to most prominent PSMs are highlighted. Structural motifs themselves are shown in Table 1 (the numbers are following the zone numeration in Fig. 3). As results of the analysis reveal, many data sets focus on chloroquine analogues (4-aminoquinolines), which is not surprising, as chloroquine was one of the most efficient antimalarial drugs on the market. Smaller sets focus also on specific compounds series, such as artemisinin analogues: despite being the most used antimalarial nowadays, the family is present in only three data sets (PS5, PS7 and PS10), with PS7 being almost exclusive to this series. Another example is PS4, which contains a unique quinolone-based motif. Such set- and zone-specific patterns lead to distinction of mode-of-action-specific zones: since all artemisinin-like compounds are located in a small area, this area corresponds to the same mechanism as artemisinin itself (which is believed to be oxidative metabolic pathway). In such a

**Table 1** Top privileged antimalarial structural motifs resulted from the analysis of responsibility patterns on the general map (17 data sets combined)




Numbers in parentheses correspond to the number of the zone on Fig. 3

way, the map may be able to not only provide prediction about the activity of a compound projected onto it, but also its mode of action.

Certain PSMs are more prominent in bigger data sets, studying cytotoxicity of large, diverse sets: for example, naphthoquinone, flavone and quinolone analogues are privileged here, while are almost never represented in family-specific studies. Often, their mechanism is not clear, since they are measured in in vitro parasite proliferation and large biochemical target screening assays, and often selected as false positives in high-throughput screens. Their unspecific interactions with numerous biological targets have lead to coining the concept of PAINS (pan-assay interference non-specific compounds) [42]. Besides the example of atovaquone, a licensed antimalarial 2-hydroxy-1,4-napthtoquinone derivative (marketed as malarone®, a fixed drug combination of atovaquone and proguanil), which specifically targets the mitochondrial electron chain transfer at the *bc*1 complex of *P. falciparum* parasites, certain menadione representatives are extensively studied in literature [26, 43, 44]. Many other 1,4-naphthoquinones are represented mostly in big, non-target and non-family specific data sets, with a MoA still under debate, and they are not present in smaller, scaffold-centric sets other than PS10.

Eventually, such a general representation of the data sets on the map provides insight into the chemical space of families or series of studied antimalarial compounds. As one notices, most of the nodes populated by active compounds are clustered together or are surrounded by nodes of the inactive class. Such clusters represent successful hit-to-lead optimization stories: a novel series of compounds is found to be active against malaria, and the chemical space is extensively investigated around that family. As an example, the center of the map (zone **1** in Fig. 3) is populated by derivatives of 4-aminoquinoline, widely studied in antimalarial tests, and represent what may be called 'explored territory' in the chemical space of antimalarial compounds.

However, some active nodes on the map are isolated from others. These nodes are populated by small, chemically distinct series of compounds found in large screening collections, and constitute potential starting point for further investigation. For example, nodes **9** and **10** on the map (Fig. 3) contain exclusively compounds from St. Jude screening set [45] (CHEMBL730080). Such isolated

'islands' constitute '*terra incognita*' of the available chemical space: first discoveries in these zones of chemical space are made, a hit is found, but the environment is yet to be explored, and leaves many possibilities for lead optimization around that island.

## Mechanism of action hypotheses

The hand-picked set of 36 known antimalarial drugs is used to validate the map and propose a mechanism of action hypotheses. The distribution of these molecules is shown in Fig. 4. While most are projected into zones populated by active compounds, there are ten drugs that are located either in inactive nodes, or in blank areas. For example, azithromycin (**M** in Fig. 4) and mirincamycin (**N**) [46], two widely used antibiotics, are found in low-density areas, which would suggest that they are outside of the applicability domain of the map. Indeed, both azithromycin and mirincamycin are big sugar-based molecules, quite different from compounds used in the frame set and in the key data sets. Another interesting example is the case of proguanil and cycloguanil (zone **I**). Proguanil is inactive against *Plasmodium* itself, but it is metabolized in the organism into the active cycloguanil [47], so it is, per se, predicted inactive or positioned outside of the AD of a model built on functional assay data. Cycloguanil, on the other hand, is projected close to its active analogues found in some data sets, although is not in the same node due to slight differences in the chemical structure.

As another example, menoctone and atovaquone (zone **G**) are both naphthoquinone drugs, but menoctone is located in the 'inactive' area. This is explained, however, by the structural differences between these compounds: menoctone (an analogue of ubiquinone) has a long aliphatic carbon chain, and no active compound in PS10 (the only data set that contains similar naphthoquinones) has that structural feature.

Additionally, molecules of this external set were annotated with a putative or confirmed mode of action. Their distribution on the map would therefore allow one to make a MoA hypothesis for the compounds projected into the same areas due to the neighborhood principle. Four types of mechanisms are considered: inhibition of hemozoin formation, folate synthesis inhibition, redox-active related MoA, and others, including general antibiotic effect or not yet established. As one can see, zone **C** at the center of the map is dominated by drugs inhibiting hemozoin formation (Fig. 4). This is not surprising since this zone is mostly populated by 4-aminoquinolines (chloroquine analogues) for which this mode of action is confirmed experimentally [48]. Similar mode of action can be supposed for the compounds which populate the top left corner (**A** and **B**) in vicinity of projections of drugs of artemisinin family (arteflene, artemotil, etc.). Redox-based drugs occupy zones **G** and **H**. Thus, zone **G** is populated by naphthoquinones [49] inhibiting mitochondrial electron transport chain, whereas zone **H** includes methylene blue and primaquine inducing the oxidative stress in the blood stage of the parasite [50, 51]. Folic acid synthesis inhibitors [52] are more dispatched over the map, which could be explained by both their structural diversity and variety of their biological target. Certain compounds, such as pyrimethamine in **J**, affect directly dihydrofolate reductase, while others (e.g. sulfalene, sulfadoxine, dapsone, in **K** and **L**) inhibit dihydropteroate synthase. Other molecules are mostly general use antibiotics, and thus act through their respective pathways, not specific to malaria parasite.

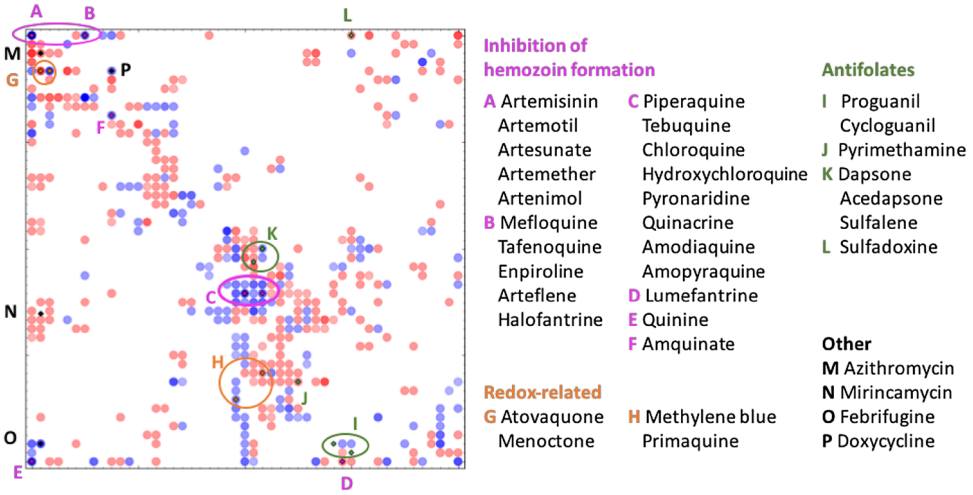



**Fig. 4** Distribution of 36 known antimalarial drugs on the map. Four different types of mechanisms are considered (see text). *Highlighted zones* (*A*–*P*) assemble molecules acting according to particular mechanism. Notice that one node may accommodate several drugs. Color code for nodes corresponds to that in Fig. 4

## Conclusions

The key merit of the paper is due to the high polyvalence of Generative Topographic Mapping, which is both, a tool for visualization, clustering and rationalization of chemical information, and a tool for predictive modeling (and as such was compared to state-of-the-art Support Vector Machine method). None of these issues are new per se, but the net benefit stems from them being reunited within a same framework, i.e. in GTM. Predictive modeling is the rigorous and quantitative branch of chemoinformatics. It can be easily benchmarked as model quality can be judged on an absolute scale, on the basis of quantitative criteria. Clustering, visualization and other procedures aimed at rationalization and intuitive rendering of chemical information are, by contrast, intrinsically subjective. There are very many distinct clustering and chemical space mapping techniques, as mentioned in introductory part. Their outputs may significantly diverge, but it is not easy to claim whether some of the methods are "better" or more "meaningful" than the other. The strength of GTM methodology is that one may accept the robustness of the predictive models of the GTM as an indirect, but solid proof of the intrinsic "meaningfulness" of the proposed visualization and clustering.

In the difficult field of antimalarial activity, where mechanistic information is sparse, while primary testing results stem from various, not always directly comparable testing protocols, GTMs provide a unique opportunity to bring together all this partial, complementary data into a common conceptual framework. It has been shown that such a common conceptual framework—represented by the consensus GTM manifold constructed on the basis of best suited "consensus" molecular descriptors—does exist. Its predictive power, with respect to all the various antimalarial activity measures, remained close to the independent modeling, in a previous work, of these same activities—including the liberty to select dedicated descriptors best suited for each activity. Also, the GTM technology provides several degrees of freedom controlling its clustering behavior. It may, for example, mimic clustering results obtained by Kohonen maps by simply decreasing the fuzziness control parameters (Gaussian weights), because GTMs are a probabilistic generalization of Kohonen maps. Herewith, GTMs maximize the chances to produce meaningful clustering schemes, and provide the quantitative criteria to support the obtained layouts.

Based on this solid ground, we highlighted some of the specific privileged patterns linked to antimalarial activity, as suggested by these maps. Of course, there are many other ways to address this question, but the one adopted herein has the merit of being backed by robust predictive statistics of relevant antimalarial activities. Therefore, the maps *might* be able to shed new light on mechanistic issues—that compounds clustered together by this approach might act through the same mechanism is a meaningful working hypothesis, which requires extensive experimental testing outside of the competence domain of our groups.

Concretely, we have demonstrated the efficiency of the GTM approach to visualize and interpret large dataset of 3000 molecules possessing antimalarial activity, to build predictive models for antimalarial activity and to make some suggestions concerning the mode of action of studied compounds.

Two dimensional GTM map well separates active and inactive compounds in distinct clusters populated by molecules with particular chemotypes. The data distribution is visualized in form of Activity Class Maps providing with an overview of data coverage of the map with density-dependent color intensity modulation. Such a representation is intuitive for medicinal chemists: zones populated by active or inactive compounds are easily identified and may guide chemist in the search of the important structural features in the lead optimization process. Additionally, isolated 'islands' populated by actives represent structurally novel hits found in large screens, which would be interesting starting points for more detailed SAR studies.

17 "local" GTM-based classification models were derived from the respective key sets corresponding to different anti-*Plasmodium* activity assessment protocols. They perform similarly to previously reported SVM models in both cross-validation and external test set prediction.

GTM allowed us to analyze the data distribution through the prism of responsibility patterns, which helps to identify the most prominent structural motifs of the antimalarial agents. Finally, by projecting compounds annotated by their mode of action against *Plasmodium*, the map allowed us delineate in the chemical space several zones populated by the compounds with selected mechanism of action.

**Acknowledgements** The Laboratory of Chemoinformatics wishes to thank the High Performance Computing centers of the University of Strasbourg, France and the Babes-Bolyai University of Cluj, Romania for supplied computer power, and assistance. P.S. thanks Program of Competitive Growth of Kazan Federal University for support. B.V. is grateful to the European Social Fund (Grant 30.1-9.1/575, mediated by Archimedes Foundation, http://www.archimedes.ee, DoRa T6 subprogram, internationalization and mobility support scheme, for the mobility stipend) and the COST Action CM1307 for three short-term scientific missions (STSM) fellowships to Strasbourg to perform data analysis, curation and database creation, as well as the cheminformatics study. B.V. and U.M. are also grateful for financial support from the Estonian Ministry of Education and Research (Grant IUT34-14). This work was partly supported by the International Center for Frontier Research in Chemistry in Strasbourg (icFRC Innovation 2015 Program, project entitled "Computer-Aided Design of Novel Antimalarial Naphthoquinones (CAD-NQ)", A.V., E.D.-C.) and the Laboratoire d'Excellence ParaFrap (grant LabEx ParaFrap ANR-11-LABX-0024, E.D.-C.).

## References


1. Barnes KI (2012) Antimalarial drugs and the control and elimination of malaria. In: Staines HM, Krishna S (eds) Treatment and prevention of malaria: antimalarial drug chemistry, action and use. Springer, Basel, p 1
2. World Health Organization (2016) World malaria report 2015. World Health Organization, Geneva
3. Snow RW, Guerra CA, Noor AM, Myint HY, Hay SI (2005) Nature 434(7030):214
4. Hott A, Tucker MS, Casandra D, Kyle DE (2015) J Antimicrob Chemother 70(10):2787
5. Liu H, Ding Y, Walker LA, Doerksen RJ (2015) Chem Res Toxicol 28(2):169
6. Blank O, Davioud-Charvet E, Elhabiri M (2012) Antioxid Redox Signal 17(4):544
7. Winzeler EA (2008) Nature 455(7214):751
8. Jensen M, Mehlhorn H (2009) Parasitol Res 105(3):609
9. Flannery EL, Fidock DA, Winzeler EA (2013) J Med Chem 56(20):7761
10. Plouffe D, Brinker A, McNamara C, Henson K, Kato N, Kuhen K, Nagle A, Adrián F, Matzen JT, Anderson P, Nam T-g, Gray NS, Chatterjee A, Janes J, Yan SF, Trager R, Caldwell JS, Schultz PG, Zhou Y, Winzeler EA (2008) Proc Natl Acad Sci 105(26):9059
11. Solomon VR, Puri KS, Srivastava K, Katti BS (2005) Bioorg Med Chem 13:2157–2165
12. Gupta MK, Prabhakar YS (2006) J Chem Inf Model 46(1):93
13. Deshpande S, Solomon VR, Katti BS, Prabhakar SY (2009) J Enzyme Inhib Med Chem 24:94–104
14. Luan F, Xu X, Cordeiro MN, Liu H, Zhang X (2013) Curr Comput Aided Drug Des 9(1):95
15. Autreto PAdS, Lavarda FC (2008) Revista do Instituto de Medicina Tropical de São Paulo 50:21
16. de Campos LJ, de Melo EB (2014) J Mol Graph Model 54:19
17. Beheshti A, Pourbasheer E, Nekoei M, Vahdani S (2016) J Saudi Chem Soc 20(3):282
18. Gamo F-J, Sanz LM, Vidal J, de Cozar C, Alvarez E, Lavandera J-L, Vanderwall DE, Green DVS, Kumar V, Hasan S, Brown JR, Peishoff CE, Cardon LR, Garcia-Bustos JF (2010) Nature 465(7296):305
19. Ludin P, Woodcroft B, Ralph SA, Mäser P (2012) Int J Parasitol 2:191
20. Spitzmüller A, Mestres J (2013) PLoS Comput Biol 9(10):e1003257
21. Wawer M, Bajorath J (2011) ACS Med Chem Lett 2(3):201
22. Spangenberg T, Burrows JN, Kowalczyk P, McDonald S, Wells TNC, Willis P (2013) PLoS One 8(6):e62906
23. Gaulton A, Bellis LJ, Bento AP, Chambers J, Davies M, Hersey A, Light Y, McGlinchey S, Michalovich D, Al-Lazikani B, Overington JP (2011) Nucleic Acids Res 40(D1):D1100
24. Belorgey D, Antoine Lanfranchi D, Davioud-Charvet E (2013) Curr Pharm Des 19(14):2512
25. Elhabiri M, Sidorov P, Cesar-Rodo E, Marcou G, Lanfranchi DA, Davioud-Charvet E, Horvath D, Varnek A (2015) Chem Eur J 21(8):3415
26. Sidorov P, Desta I, Chessé M, Horvath D, Marcou G, Varnek A, Davioud-Charvet E, Elhabiri M (2016) ChemMedChem 11(12):1339
27. Kireeva N, Baskin, II, Gaspar HA, Horvath D, Marcou G, Varnek A (2012) Mol Inform 31(3–4):301
28. Viira B, Gendron T, Lanfranchi D, Cojean S, Horvath D, Marcou G, Varnek A, Maes L, Maran U, Loiseau P, Davioud-Charvet E (2016) Molecules 21(7):853
29. Horvath D, Marcou G, Varnek A (2013) J Chem Inf Model 53(7):1543
30. ChemAxon (2008) Standardizer. ChemAxon, Budapest. http://www.chemaxon.com/jchem/doc/user/standardizer.html. Accessed 20 Feb 2009
31. ChemAxon (2007) Tautomer Plugin. ChemAxon, Budapest. http://www.chemaxon.com/marvin-archive/4.1.3/marvin/chemaxon/marvin/help/calculator-plugins.html - tautomer. Accessed 20 Oct 2011
32. Ruggiu F, Marcou G, Varnek A, Horvath D (2010) Mol Inform 29(12):855
33. Varnek A, Fourches D, Horvath D, Klimchuk O, Gaudin C, Vayer P, Solov'ev V, Hoonakker F, Tetko Iv, Marcou G (2008) Curr Comput Aided Drug Des 4(3):191
34. Varnek A, Fourches D, Hoonakker F, Solov'ev VP (2005) J Comput Aided Mol Des 19(9–10):693
35. Laboratoire de Chemoinformatique Strasbourg (2012) Nomenclature of ISIDA fragments
36. Gaspar HA, Sidorov P, Horvath D, Baskin II, Marcou G, Varnek A (2016) Generative topographic mapping approach to chemical space analysis. In: Frontiers in molecular design and chemical information science—Herman Skolnik Award Symposium 2015: Jürgen Bajorath, vol 1222. American Chemical Society, p 211
37. Gaspar H, Marcou G, Horvath D, Arault A, Lozano S, Vayer P, Varnek A (2013) J Chem Inf Model 53(12):3318
38. Gaspar HA, Baskin II, Marcou G, Horvath D, Varnek A (2015) Mol Inform 34(6–7):348
39. Sidorov P, Gaspar H, Marcou G, Varnek A, Horvath D (2015) J Comput Aided Mol Des 29(12):1087
40. Horvath D, Brown BJ, Marcou G, Varnek A (2014) Challenges 5(2)
41. Klimenko K, Marcou G, Horvath D, Varnek A (2016) J Chem Inf Model 56(8):1438
42. Baell JB, Holloway GA (2010) J Med Chem 53(7):2719
43. Müller T, Johann L, Jannack B, Bruckner M, Lanfranchi DA, Bauer H, Sanchez C, Yardley V, Deregnaucourt C, Schrevel J, Lanzer M, Schirmer RH, Davioud-Charvet E (2011) J Am Chem Soc 133(30):11557
44. Lanfranchi DA, Cesar-Rodo E, Bertrand B, Huang H-H, Day L, Johann L, Elhabiri M, Becker K, Williams DL, Davioud-Charvet E (2012) Org Biomol Chem 10(31):6375
45. Guiguemde WA, Shelat AA, Bouck D, Duffy S, Crowther GJ, Davis PH, Smithson DC, Connelly M, Clark J, Zhu F, Jiménez-Díaz MB, Martinez MS, Wilson EB, Tripathi AK, Gut J, Sharlow ER, Bathurst I, El Mazouni F, Fowble JW, Forquer I, McGinley PL, Castro S, Angulo-Barturen I, Ferrer S, Rosenthal PJ, Derisi JL, Sullivan DJ, Lazo JS, Roos DS, Riscoe MK, Phillips MA, Rathod PK, Van Voorhis WC, Avery VM, Guy RK (2010) Nature 465(7296):311
46. Krishna S, Staines HM (2012) Non-antifolate antibiotics: clindamycin, doxycycline, azithromycin and fosmidomycin. In: Staines HM, Krishna S (eds) Treatment and prevention of malaria: antimalarial drug chemistry, action and use. Springer, Basel, p 141
47. Olliaro P (2001) Pharmacol Ther 89(2):207
48. O'Neill PM, Barton VE, Ward SA, Chadwick J (2012) 4-Aminoquinolines: chloroquine, amodiaquine and next-generation analogues. In: Staines HM, Krishna S (eds) Treatment and prevention of malaria: antimalarial drug chemistry, action and use. Springer, Basel, p 19
49. Vaidya AB (2012) Naphthoquinones: atovaquone, and other antimalarials targeting mitochondrial functions. In: Staines HM, Krishna S (eds) Treatment and prevention of malaria: antimalarial drug chemistry, action and use. Springer, Basel, p 127
50. Baird K (2015) Pathog Glob Health 109(3):93
51. Hill DR, Baird JK, Parise ME, Lewis LS, Ryan ET, Magill AJ (2006) Am J Trop Med Hyg 75(3):402
52. Nzila A (2012) Antifolates: pyrimethamine, proguanil, sulphadoxine and dapsone. In: Staines HM, Krishna S (eds) Treatment and prevention of malaria: antimalarial drug chemistry, action and use. Springer, Basel, p 113

# Chapter 7

# Modeling redox properties of antimalarial compounds

Among many modes of action of antimalarial compounds, one of the most prominent for known compounds is the heme detoxification during the hemoglobin digestion. The malarial parasite *P. falciparum* digests its host cell's hemoglobin as a source of essential nutrients, that leads to the formation of toxic $Fe^{III}$ ferriprotoporphyrin (or hematin). The parasite can evade the toxicity by several ways, the most prominent of which are either polymerizing the hematin into insoluble hemozoin (malaria pigment), or reducing it in a thiol network in the cytosol. Also, methemoglobin($Fe^{III}$) (metHb) is digested faster than Hb($Fe^{II}$).

The pathway of heme detoxification by heme polymerization is known to be disrupted by 4-aminoquinoline series of compounds (examples of which are chloroquine and related molecules): they bind to ferriprotoporphyrin and prevent the formation of hemozoin [21, 33]. The second mechanism considers the interference with the redox cycles of the parasite, for example, in gluthathione reductase (GR) [204]. Thus, redox active compounds (such as naphthoquinones) may slow the parasite's metHb digestion rate by reducing it and, at the same time, decrease the hemozoin formation. The supposed mechanism of action of quinone family compounds is shown on the Figure 7.1.

Redox active compounds of other families are also investigated for their antimalarial antimalarial activity, for example, indolone-N-oxide derivatives [205], and the studies confirm their participation in redox cycles of infected red blood cells. Therefore, redox properties (such as redox potential) are of great importance for directed design of antimalarial drugs. This chapter describes the QSPR models for redox potentials of naphthoquinones compounds family, and one new type of descriptors developed for this project.

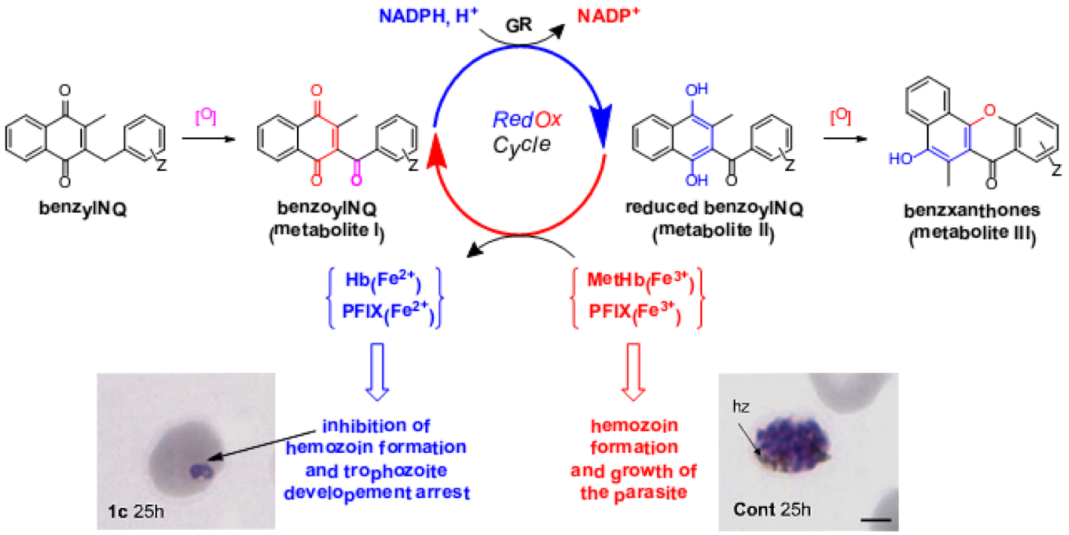



FIGURE 7.1: Supposed mechanism of antimalarial action of naphthoquinone derivatives. Benzoyl-substituted naphthoquinones interfere with the GR-related redox cycle in the parasite, thus interrupting the heme detoxification process. [201]

## 7.1 Oxidation/reduction potential

The oxidation/reduction (or redox for short) potential is a quantitative expression of the tendency of a compound to give or take electrons [206]. It may be compared to an acid-base reaction, though in the case of oxidation-reduction reaction there is the electron transfer. The redox potential corresponds to single oxidation-reduction equilibrium (half-reaction) and obeys the Nernst equation:

$$Ox \xrightleftharpoons{+ne^-} Red; E = E^0 + \frac{RT}{nF} \ln \frac{a_{Red}}{a_{Ox}} \tag{7.1}$$

where $E$ is the half-reaction reduction potential, $E^0$ is the standard half-reaction reduction potential, invariant to experimental conditions, $R$ is the universal gas constant, $T$ is the absolute temperature, $a$ is the chemical activity for the relevant species, $F$ is the Faraday constant, $n$ is the number of moles of electrons transferred in the half-reaction [207].

$E^0$, the standard redox potential, is characteristic for each oxidation-reduction pair and depends on the Gibbs free energy of the oxidation-reduction reaction [207]:

$$\Delta G^0 = -nFE^0. \tag{7.2}$$

So, it is clearly seen that the value of the redox potential depends on the free energy of the reaction, as well as the temperature.

Different techniques exist for measuring the redox potential [207]. Generally they are divided into potentiometric (based on the Nernst equation), coulometric (based on the Faraday's law) and voltammetric and amperometric (scanning specific potential scale) methods. Here, we have used data coming from cyclic voltammetry. In cyclic voltammetry the potential is scanned in both directions – for oxidation and reduction. This way, the curve has separate peaks for oxidation and reduction, and the potential of the reaction is calculated as the average of them [208].

As bibliographic research shows, the attempts to predict the redox potential (especially for this specific compounds family) are not very numerous. Firstly and evidently, the predictions based on thermodynamic functions of the formation of anion-radical as the first step of reduction of quinones catch the eye. Most of these models use Gibbs free energy of oxidized and reduced forms, calculated using either heavy Hartree-Fock [209–211] and DFT [212] quantum chemical calculations, or fast, but less precise semi-empirical methods. A QSPR analysis of redox potentials of quinones in different solvents has been also conducted [213] and provides an MLR model with variables calculated with semi-empirical methods, as well as solvent properties as Guttmann acceptor number and dipole moment.

A notable study [214] of redox potential of different quinone families (benzo-, naphtho- and anthraquinone derivatives) proposes a simple regression model, based on specific quantum chemical indices [215]. Though the initial dataset was not vast (26 molecules), the resulting equation, including only one variable (electrophilicity index), is reported to be robust when challenged against external tests.

Still, almost all previous predictions use descriptors based on costly quantum chemical calculations (DFT or Hartree-Fock). This could make the search for a compound having specific redox properties, using combinatorial libraries and virtual screening, difficult, so the goal of finding a simpler model is still not achieved.

## 7.2 Electronic Effect descriptors

Electronic Effect Descriptors (EED) represent a generalization of a previously advocated attempt to empirically characterize the global impact of the chemical environment on the electron density of a given "key" atom $K$ in a molecule, by descriptors designed as simple additive topological distance-weighted contributions of each of the $N$ atoms of the compound. In the present, EEDs were defined as:

$$EED_{p,e,o,w,c}(K) = \sum_{i=1}^{N} \delta^c(i,K) \times P_p^e(i) \times \exp\left(-\frac{(\tau_{iK} - o)^2}{w}\right) \tag{7.3}$$

Above, the five indices $p, e, o, w, c$ go over all the considered combinations of options, as outlined below:

- $p = 1...11$ browses through the set of considered atomic properties $P_p$. Properties associated to each $p$ values are given in Table 7.1.
- $e = 1...2$ stands for the exponent applied to the property values – either plain values, or their squares are alternatively considered.
- $o = 1...4$ and $w = 2...5$ control the offset and the width of the Gaussian functions of the shortest path topological distance used to encode the participation of atom $i$ to the effect on key atom $K$. If $o = 0$, the self-term of the key atom $K$ would be the major contributor, whilst atoms of successive coordination shells contribute less ($w$ controls the contribution decay rate). At $o > 0$, the descriptor terms single out atoms within coordination shell #$o$ as main contributors, whilst atoms both closer and further to $K$ contribute less. For convenience (in order to be able to eventually explore alternative functional dependencies having singularities at zero topological distance, $i = K$), the variable $\tau_{iK}$ in equation 7.4 was chosen to represent topological distances in all cases except when $i = K$ (see below). With this small trick, calculating descriptors at both $o = 0$ and $o = 1$ is no longer needed – the scanned $o$ range starts at 1, as outlined above.

$$\tau_{iK} = \begin{cases} 0.5 & \text{if } i = K \\ \text{shortest path}(i - K) & \text{otherwise} \end{cases} \tag{7.4}$$

- $c = 0, 1$ is a conjugation control parameter, formally acting through the switching function $\delta^c(i,K)$ defined in equation 7.5 below. Therefore, all descriptor elements associated to $c = 0$ represent sums over all the $N$ atoms in a molecule, whereas at $c = 1$ only atoms $i$ connected to $K$ by means of a path including only unsaturated atoms (except for $i$ and $K$ themselves) are considered. "Unsaturated" status is checked by means of ChemAxon hybridization flags – none of the atoms between $i$ and $K$ (exclusive) should be flagged $sp^3$, or else the contribution of $i$ is ignored in terms associated to $c = 1$.

$$\delta^c(i,K) = \begin{cases} 1 & \text{if } c = 0 \text{ or path between } i \text{ and } K \\ & \text{contains only unsaturated atoms} \\ 0 & \text{otherwise} \end{cases} \tag{7.5}$$

| $p$ | Label | Atomic Property $P_p$ | Source and remarks |
|---|---|---|---|
| 1 | CHG | Partial Charge | Gasteiger-Marsili [216] charge – ChemAxon charge plugin [217] |
| 2 | CHS | Sigma Charge | Gasteiger-Marsili [216] charge – ChemAxon charge plugin [217] |
| 3 | CHP | Pi Charge | Gasteiger-Marsili [216] charge – ChemAxon charge plugin [217] |
| 4 | POL | Polarizability | ChemAxon polarizability plugin [218] |
| 5 | ELN | Relative Electronegativity with respect to carbon | Pauling electronegativities from ChemAxon API |
| 6 | ENU | Nucleophilic Energy Index | ChemAxon Hückel Analysis plugin [219] |
| 7 | EEL | Electrophilic Energy Index | ChemAxon Hückel Analysis plugin [219] |
| 8 | EDI | Electron Density Index | ChemAxon Hückel Analysis plugin [219] |
| 9 | CDI | Charge Density Index | ChemAxon Hückel Analysis plugin [219] |
| 10 | HYB | Hybridization Index | ChemAxon API |
| 11 | FCG | Formal Charge | ChemAxon API (standardization[166]-dependent) |

Table 7.1: Atomic properties entering the calculation of EEDs.

The $11 \times 2 \times 4 \times 4 \times 2 = 704$ combinations of control parameters define a 704-dimensional EED vector, the elements of which are labeled {PropertyLabel}^$e$.off$o$.w$w$.conj$c$. So, CHG^2.off1.w2.conj0 represents the summation of squares of partial charges, centered on the chosen key atom of the molecule, using a topological distance-dependent functional form offset by 1 and with a width parameter of 2, and irrespective of the saturation status of interposed atoms.

## 7.3 Redox potential modeling

### 7.3.1 Dataset and descriptors

Our colleagues from the Laboratory of Bioorganic and Medicinal Chemistry (University of Strasbourg, France) have provided us with a dataset of substituted naphthoquinones with the redox potential measured in cycic voltammetry. In order to enrich the chemical space of the set, other quinones (various benzo-, naphtho-, anthraquinones) [214] and indolone-N-oxides (potential antimalarial agents also acting *via* oxidative stress) [205] from the literature have been added. The equivalence of experimental measurement conditions and the modeled property (first wave redox potential) was established. Where the redox potential values were measured with a different reference electrode, they were accordingly offset by the reference electrode potential difference prior to their merging with in-house data. Variations due to slightly different solvent and temperature

in measurement protocols were not significant and thus were neglected [128]. In total, we selected 95 compounds measured under similar experimental conditions.

The structures were standardized (aromatization and charge-separated representation of nitro-groups) prior to modeling using the ChemAxon Standardizer tool [166], version 14.10.20 (http://www.chemaxon.org).

Two types of descriptors were used: Electronic Effect Descriptors (EED) and ISIDA fragment descriptors [126]. The generation of EEDs is described above. Here, the carbon atom in the carboxyl group of the quinone or indolone moiety was chosen as a key atom. In the case of the quinone ring, two carboxyl groups are present, so the descriptors are calculated as an average value between the two.

ISIDA fragments offer several advantages over other types of descriptors: versatility, simplicity and rapidity of calculations, various applicability domain schemes [220, 221] (fragment control and bounding box strategies). With that in mind, 44 different fragmentation schemes, representing atom and bond sequences in different lengths ranging from 2 to 10, were used for training and building a consensus model.

### 7.3.2 Computational procedure

ISIDA/QSPR software [222] was used to build multilinear regression models [223] on the training set consisting of 81 molecules. The external test set contained 14 compounds that were either voluntarily kept aside or actually measured after model fitting. The software offers consensus model building with any number of fragmentation schemes, as well as variable selection and applicability domain monitoring during the model selection process.

In addition, nonlinear models (SVM epsilon-regression) were also created, using the evolutionary libsvm tuning tool [190] for optimization. This ensures optimal selection of the descriptor space (of both ISIDA and EED descriptor candidates) and of SVM parameters yielding maximal robustness (*i.e.*, best cross-validation performance) models. After optimization, EED descriptors were selected as the most powerful and were used henceforth in the SVM model.

### 7.3.3 Results and discussion

The models were built and evaluated by both methods following a three-fold cross-validation procedure repeated ten times to reduce the influence of random factor and

were additionally evaluated on the designated external set. For ISIDA descriptors, a consensus model was applied with ISIDA/QSPR's FMF module. The top EED-based SVM model was deployed on the web server application (`http://infochim.u-strasbg.fr/webserv/VSEngine.html`) and then used for external validation.

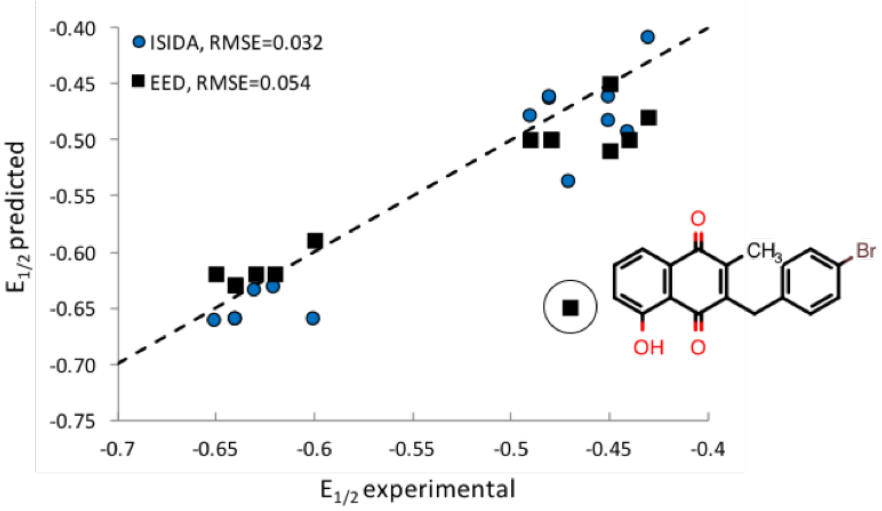



FIGURE 7.2: Plot of $E^1_{1/2}$ for test set compounds, experimental values vs. those predicted by all models: ISIDA MLR (•), predicted by ISIDA/QSPR software, and EED SVM (■), predicted by the web server application. The dotted line represents perfect fit. The outlier (benzyl derivative with hydroxy substituent in the naphthoquinone moiety) is highlighted.

The results are presented in detail further in the articles. Table 7.2 reports briefly the statistical parameters for the models built with two methods. The cross-validation $R^2$ scores were quite high for all modeling methods. The consensus models based on ISIDA descriptors also performed well in case of external test set, with the $R^2$ of the test being close to the $R^2$ of cross-validation. The model with EED descriptors had a globally lower $R^2$, but this was due to one molecule being predicted more poorly, which, due to the modest size of the test set, had a greater impact on the overall score (Figure 7.2).

The poorly predicted molecule is the derivative with a hydroxy substituent, and the phenomenon has been reported previously [128]. However, the training set in this work

| Model | 3-fold CV | | Test | |
|---|---|---|---|---|
| | *RMSE* | $R^2$ | *RMSE* | $R^2$ |
| ISIDA MLR | 0.080 | 0.85 | 0.032 (0.027)* | 0.86 (0.90)* |
| EED SVM | 0.082 | 0.84 | 0.054 (0.030)* | 0.60 (0.88)* |

TABLE 7.2: Statistical characteristics of consensus multilinear regression (ISIDA MLR) and support vector machine (EED SVM) models for cross-validation (CV) and external test challenges. *Adjusted statistical parameters when the outlier is excluded from the test set.

includes examples of analogues with such an intramolecular hydrogen bond and, accordingly, the structural pattern was recognized when using ISIDA fragment counts. The model learned to associate this fragment with a decrement for the predicted redox potential (the hydrogen bond polarizes the quinone carbonyl even further, rendering the quinone system more electron-depleted). The presence of such a fragment itself is not sufficient to trigger a decrease in the redox potential; it must be determined whether the fragment is in the right context, that is, involving the actual quinone carbonyl and the adjacent phenolic hydroxy group in a hydrogen bonding position. Should the mentioned sequence appear in other moieties of a molecule, not connected to the quinone system, the model would also add the associated (negative) increment to the predicted redox potential and would likely result in an error. This shows that even an extended compound set used in this work is still prone to significant biases and sources of artifacts.

In contrast, EED descriptors do not explicitly consider connectivity patterns but only focus on through-bond electronic effects, ignoring all through-space non-bonded influences. As such, they cannot capture the atypical effect in 5-OH analogues, even if these were present in the training set. They fail to learn any specific rule for the 5-OH analogues and therefore commit an important error of +0.17 V when predicting the redox potential of the outlier. However, they are less prone to fail in the scenario of external predictions for compounds with similar moieties not responsible for redox behavior. If the molecule is discarded as an outlier, then $RMSE = 0.030$ V and $R^2 = 0.88$ (comparatively, for ISIDA descriptors $RMSE = 0.027$, $R^2 = 0.90$).

## 7.4 Conclusion

Two different descriptor types – conventional ISIDA SMF and newly developed EED – were used to build models to predict the first wave redox potential $E^1$ of naphthoquinones and similar compounds. The developed QSPR models were successful in modeling the redox potential applicable to several different redox-active scaffolds — benzo-, naphtho-, antraquinones and indolone-N-oxides. They succeeded in both cross-validation and external test prediction challenge. These predictions were quantitatively accurate, experimental and predicted values being within the nominal accuracy range of the models.

The models have been deployed on the web server application on our web-site (`http://infochim.u-strasbg.fr/webserv/VSEngine.html`) and are accessible for public use. Evidently, the applicability domain of the model accepts only molecules with a carboxyl redox center, preferably a quinone or indolone moiety. Performance with other scaffolds of redox-active compounds is not guaranteed.

ChemPubSoc Europe

DOI: 10.1002/chem.201403703

CHEMISTRY A European Journal Full Paper

Redox Chemistry

# Electrochemical Properties of Substituted 2-Methyl-1,4-Naphthoquinones: Redox Behavior Predictions

Mourad Elhabiri,[a] Pavel Sidorov,[b, c] Elena Cesar-Rodo,[a] Gilles Marcou,[b] Don Antoine Lanfranchi,[a] Elisabeth Davioud-Charvet,*[a] Dragos Horvath,[b] and Alexandre Varnek*[b, c]

**Abstract:** In the context of the investigation of drug-induced oxidative stress in parasitic cells, electrochemical properties of a focused library of polysubstituted menadione derivatives were studied by cyclic voltammetry. These values were used, together with compatible measurements from literature (quinones and related compounds), to build and evaluate a predictive structure–redox potential model (quantitative structure–property relationship, QSPR). Able to provide an online evaluation (through Web interface) of the oxidant character of quinones, the model is aimed to help chemists targeting their synthetic efforts towards analogues of desired redox properties

## Introduction

The quinone structure, which is common to numerous natural products with important biological activities, is known for its ability to accept one and/or two electrons in redox processes.[1] The electron-acceptor properties of quinones, causing the formation of radical semiquinone anion or dihydroquinone dianion species responsible for in vivo oxidative stress,[2] can be modulated by the electron-withdrawing or -donating substituents of the electroactive core. The molecular basis of quinone toxicity is the enzyme-catalyzed reduction of the quinone to semiquinone radicals, which then reduce $O_2$ to superoxide anion radicals and hydrogen peroxide through 1 e- or 2 e-transfer reactions thereby regenerating the quinone. This futile redox cycling and concomitant oxygen activation leads to increased levels of reactive oxygen species (ROS) and glutathione disulfide.[3,4] A well-known example is menadione (2-methyl-1,4-naphthoquinone or vitamin K3), which is a redox-cycler or a "subversive substrate" for numerous flavoproteins acting through a one-electron reduction mechanism, for example, the nicotinamide adenine dinucleotide phosphate (NADPH)-dependent glutathione reductase,[5,6] the NAD(P)H dehydrogenase, lipoamide dehydrogenase (LipDH),[7,8] the trypanothione reductase,[7–9] the thioredoxin reductase[5,10] or the thioredoxin-glutathione reductase.[11] Anecdotally, lipoamide dehydrogenase was named menadione reductase in earlier times.

Menadione and its 5-hydroxylated analogue (plumbagin), are important examples of the broad family of 1,4-naphthoquinones (1,4-NQs), largely distributed in nature (Figure 1). Menadione is the parent core of vitamins K1 and K2. Vitamin K1 (phylloquinone, phytomenadione, or phytonadione) is only

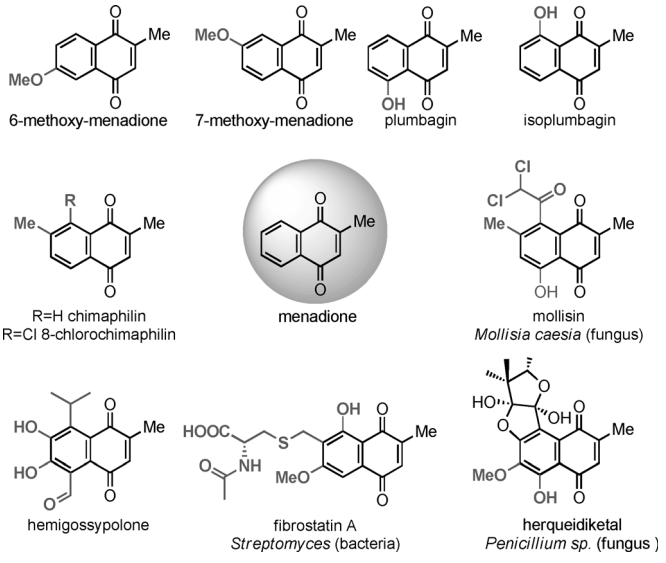



**Figure 1.** Natural menadione derivatives polysubstituted at the aromatic ring including menadione (2-methyl-1,4-naphthoquinone) and plumbagin (5-hydroxy-menadione).

[a] Dr. M. Elhabiri, E. Cesar-Rodo, Dr. D. A. Lanfranchi, Dr. E. Davioud-Charvet
Laboratoire de Chimie Bioorganique et Medicinale
UMR7509 CNRS-Université de Strasbourg
European School of Chemistry, Polymers and Materials (ECPM)
25 Rue Becquerel, F-67087 Strasbourg (France)
Fax: (+33) 3-68-85-27-42
E-mail: elisabeth.davioud@unistra.fr

[b] P. Sidorov, Dr. G. Marcou, Dr. D. Horvath, Prof. A. Varnek
Laboratoire de Chemoinformatique
UMR 7140 CNRS-Université de Strasbourg
1 rue Blaise Pascal, Strasbourg 67000 (France)
E-mail: varnek@unistra.fr

[c] P. Sidorov, Prof. A. Varnek
Butlerov Institute of Chemistry
Kazan Federal University, Kazan (Russia)

Supporting information for this article is available on the WWW under http://dx.doi.org/10.1002/chem.201403703.

produced in plants; the menadione core is alkylated by a phytyl chain at the C3 position. Vitamin K2 (menaquinone) represents a series of compounds in which the phytyl side chain of phytonadione has been replaced by a side chain built up of 1 to 14 isoprenyl units. In mammals, reduced vitamin K2 (as opposed to the oxidized form, vitamin K 2,3-epoxide) is the critical nutrient required for blood coagulation as a cofactor for the γ-carboxylation of glutamic acid residues in the blood-clotting protein prothrombin. It also plays a key role in bone homeostasis and is a clinically effective therapeutic agent for osteoporosis. Owing to its electron-acceptor properties, vitamin K2 has recently been shown to act as a mitochondrial electron carrier that rescues Pink1 deficiency, due to genetic mutations causing Parkinson's disease in humans and mitochondrial defects in model organisms.[12] In folk medicine, plumbagin is the main principle from powdered roots of the plant, *Plumbago zeylanica*. It is broadly administered as home remedies for its antileishmanial activities upon different *Leishmania* spp. responsible for human parasitic diseases that include cutaneous and mucocutaneous forms of leishmaniasis.[13] Menadione derivatives display a broad pattern of diverse biological responses, for example, antibacterial, anticancer, antifungal, antitrypanosomal,[8,14] antiinflammatory, antimalarial,[15–17] or anti-Alzheimer activities, to cite a few. It is therefore not surprising to notice that these redox scaffolds have been employed as valuable molecular templates in drug discovery. Because the 1,4-naphthoquinone motif is broadly represented in nature, there is a growing renewal of interest in medicinal chemistry projects to prepare polysubstituted 1,4-NQs with peculiar redox properties (i.e., redox potentials lying within a determined redox potential range to interact with targeted redox enzymes).[18]

Our team has pioneered the biological chemistry and the molecular understanding of these redox-active agents catalyzing various NADPH-consuming reactions, which are behind the antiparasitic activities.[8,15,17,19] We recently discovered a series of potent antimalarial 3-benzyl-menadiones,[15] which require a bioactivation step through redox domino reactions catalyzed by both glutathione reductases of *Plasmodium*-infected red blood cells. Our approach to generate oxidative stress specifically in parasitized human erythrocytes has been validated as a new and efficient strategy to combat malarial parasites.[15] Antimalarial drug development and fine-tuned optimization however necessitate fundamental approaches to assess the substitution effects on various parameters such as redox behavior, activity, metabolism, and bioavailability, to cite a few. Numerous routes to prepare menadione or *aza*-menadiones derivatives alkylated at the quinone part of the 1,4-naphthoquinone were reported.[8,14,19] However, there are not many versatile methods for the regioselective preparations of synthetic naphthoquinone derivatives bearing a methyl group at C2 of the quinone moiety (east) and substituted at the phenyl ring (west part) due to the dissymmetry of the (*aza*-)menadione molecule with its 2-methyl group.[20–22] For instance, the various combinations of the factors governing the regioselectivity of cycloadditions can produce highly selective and reactive effects; however, the results of these approaches depend on the structural features of the substrates, rendering the outcome of the reaction often unpredictable.

As synthesis resources are limited, whereas the relevant chemical space of feasible analogues is practically infinite, selection of the envisaged synthetic products was a matter of the medicinal chemist's know-how, and serendipity. As an appealing alternative, computer-aided molecule design based on quantitative structure–property relationships (QSPR)[23] is nowadays routinely used to prioritize the synthesis and testing of molecules of predicted properties, out of the virtually infinite space of possible organic structures. Successful predictive models,[24–26] able to effectively orientate synthetic chemists towards structures effectively fulfilling their demands, need first to be "learned" from existing experimental input[27] (structures represented by molecular descriptors[28–31] and associated property values). Such machine learning[32,33] amounts to a calibration of a mathematical model estimating the experimental property $P$ for a molecule $M$ as a function of relevant molecular descriptors $D_1$, $D_2$,…$D_n$. The simplest functional form, Equation (1), is linear; $c_i$ values are fitted by multilinear regression (MLR).

$$P(M) = f[D_1(M), D_2(M), ..., D_n(M)] = c_0 + \sum_{i=1}^{n} c_i D_i(M) \qquad (1)$$

Approaches based on the above philosophy are already familiar to chemists. Hammett Equations[34–36] are used to relate the impact of a substituent at a predefined scaffold position to the property shift ($pK_a$, reaction rate, etc.) of another group in the molecule. The role of the "descriptor" is played by the σ "substituent constant", to be multiplied by the sensitivity term ρ of the studied reaction. Tabulated σ values were derived on the basis of past experiments, typically for substituents on aromatic rings. In chemoinformatics, descriptors are empirical quantities calculated on the basis of structural aspects. They are no longer linked to specific substitution patterns on predefined scaffolds and are hence much more general.

So far, chemoinformatics-driven attempts to predict the redox potential are extremely scarce. Predictions of thermodynamic functions of the formation of the radical anion intermediate, as the first step of reduction of quinones, catch the eye. Most of these models use Gibbs free energy of oxidized and reduced forms, calculated by using either Hartree–Fock[37–39] and DFT[40] calculations, or faster, but less precise semi-empirical methods. A QSPR analysis of redox potentials of quinones in different solvents provided an MLR model[41] based on quantum chemical (QC) descriptors and solvent properties (Guttmann acceptor number, dipole moment). The electrophilicity index,[42] also obtained from QC calculations, was shown to robustly explain[43] the redox potentials of 26 benzo-, naphtho-, and anthraquinones. Yet, the emphasis on QC descriptors limits the size of virtual libraries screenable by such approaches.

To test the relevance of QSPR analysis for the design of antiparasitic redox cyclers (i.e., 3-benzyl-menadiones targeting NADPH-consuming glutathione reductases), we herein de-

scribed the electrochemical and absorption properties of a homogeneous series of compounds containing the menadione electrophore as a basis for QSPR model building. Resulting models were challenged to predict the redox potential of new analogues, which were further synthesized and subjected to electrochemical measurement, to verify the relevance of the predictions.

It is noteworthy that the mono- or disubstitution of menadione by a same group on its west part leads to 10 structures, whereas the substitution by two different groups increases the number of combinations to 12. Therefore, a platform of synthetic methodologies has been successfully established here, providing access to a focused library of 19 mono- or disubstituted menadione derivatives functionalized at the benzene ring (west part) of the menadione core (Figure 2).

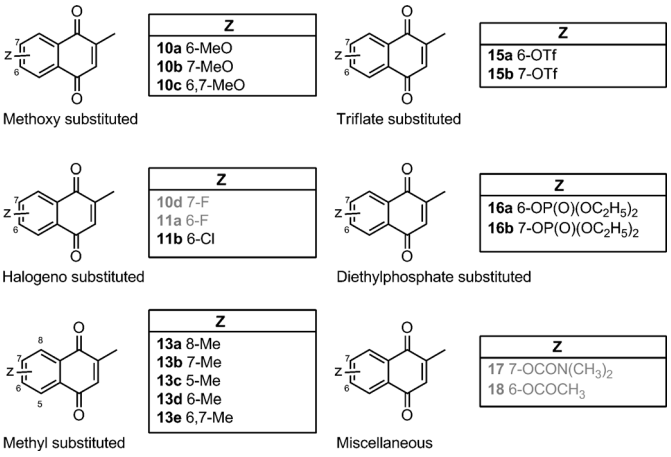



**Figure 2.** Structures of substituted menadiones. The gray color indicates the PRED series (used to test the models).

This versatile synthetic approach will be described elsewhere.[44] References menadione and plumbagin were also included in the series.

Absorption spectrophotometry and/or electrochemistry (polarography and voltammetry)[45] are informative methods to evaluate the influence of structural effects on the reactivity.[1,46] Electronic effects induced by substituents of the quinone core might cause marked shifts in the half-wave potentials, or affect the mechanism, reversibility, and/or rate of the electrode processes. It has been observed that when the substituent is directly grafted on the quinone moiety[47] or has a resonance effect on the quinone system, these significantly alter its redox properties. Analogues of the type quinone-X-substituted phenyl (X = heteroatom) have been shown[48] to finely modulate the electronic properties of the quinone backbone. Besides, inter- and intra-molecular hydrogen bonds also fine-tune the half-wave potential values. If some approaches have been focused on evaluating the substitution effects on the quinone moiety (east part) of 1,4-NQs,[49–51] very little is known about the influence of such substituents on the benzene ring[52] (west part).

A key goal of the paper is to exploit the new redox potential data, and enrich already published data, to generate simple, robust and publicly available QSPR models. Models were trained using stepwise multilinear regression: Two were based on classical descriptors (CODESSA,[31,53] including quantum-chemical (QC) terms, and ISIDA[24,25,54,55] fragment counts). Also, new, on-purpose Electronic Effect Descriptors (EED) were developed: a generalization of a previous attempt[57] to quantify inductive and resonance effect strengths based solely on the molecular connectivity. In parallel, the Hammett approach has been applied, for benchmarking purposes, to some structures for which substituent constants could be found.

Eventually, synthesis and testing of four new molecules (**10d**, **11a**, **17**, and **18**, Figure 2), completed in parallel to model development, allowed a real-life prospective prediction challenge of their redox potentials, by using above-mentioned global models.

## Results and Discussion

### Methods

This section provides an overview of the data management and modeling strategies used (see more details in the Supporting Information). A first subset of the herein used collection is formed by the 17 menadione derivatives synthesized[48] and studied (12 of which are new entities, Figure 2). These were supplemented by a series of 26 quinones with redox potential values reported in literature.[44] Since these values were measured with a different reference electrode, they were accordingly offset by the reference electrode potential difference prior to the merger with in-house data. Differences due to slightly different solvent and temperature in measurement protocols were neglected. Menadione also figured amongst the cited literature, and the therein reported values matched in-house measurements, after applying the mentioned offset. The literature set also includes benzo- and anthraquinones, providing thus a welcome increase of chemical diversity. In addition, a series of 27 redox-active indolone-*N*-oxides[58] with potentials measured under conditions similar to those used in this work, were added to the study, to expand model applicability beyond the quinone family.

Models were trained on the fused set of in-house measured and literature data. Given the quite modest set size, step-wise multilinear[59–61] regression (MLR) was used. ISIDA fragment counts were used in conjunction with the ISIDA MLR tool,[24] automatically scanning through a large pool of possible molecular fragmentation schemes to return a consensus model based on best equations found with best suited fragmentation schemes. CODESSA and EED descriptors (see the Supporting Information for details) were employed in conjunction with a stepwise MLR tool conceived as a "Lamarckian" strategy within the genetic-algorithm driven Stochastic QSAR Sampler (SQS[58]). Models were built following a 10 times repeated three-fold cross-validation[62–64] procedure. They represent consensus equations of the best "partial" equations generated during cross-validation (also see "cross-validation" in the Exper-

imental Section in the Supporting Information). To the purpose of prospective predictions of entities not yet synthesized or tested, the ISIDA fragment count-based model (a compromise maximizing both statistical robustness and user-friendliness) was deployed online,[65] allowing access to predictions through any classical browser. Four new menadione derivatives (**10d**, **11a**, **17**, and **18**, Figure 2), hereafter labeled "PRED", were used as a modest but important predictive challenge. Below, a discussion of experimental observations will be followed by modeling results.

### Absorption properties

We first recorded the electronic absorption spectra of the menadione derivatives to assess the electronic effects of the substitution on the benzene part of the 1,4-naphthoquinone (1,4-NQ) (Table 1 and the Supporting Information, Table S2 and Figures S1–S3). The low energy electronic transitions (314–351 nm) were attributed to π–π* transitions centered on the 1,4-NQ chromophore (benzoquinoidal structure). For the hydroxylated (plumbagin) and the methoxylated (**10a**, **10b**, **10c**) menadiones, an additional absorption of weaker energy (380–425 nm) was ascribed to intramolecular charge-transfer process (ICT) from the $CH_3O$- (or HO-) functions to the 1,4-NQ core.[66–69] It is noteworthy that the π–π* (benzene subunit) or n–π* transitions (quinone carbonyls or substituent heteroatoms) are either characterized by low intensities, masked by other intense transitions (π–π*(1,4-NQ), ICT) or positioned at higher energies (DMSO cut-off) and cannot be accurately observed.[70] Besides, regardless of the substitution of the menadione core, no intermolecular associations have been evidenced in solution.

**Table 1.** Spectrophotometric properties ($\lambda_{max}$, $\varepsilon^{\lambda max}$) of the substituted menadione derivatives examined in this work.[a]

| Compound | $\lambda_{max}$ ($\varepsilon^{\lambda max}$) [nm ($\times 10^3$ M$^{-1}$cm$^{-1}$)] | Compound | $\lambda_{max}$ ($\varepsilon^{\lambda max}$) [nm ($\times 10^3$ M$^{-1}$cm$^{-1}$)] |
|---|---|---|---|
| Menadione | 331(2.94) | Hydroxyl substituted | |
| Methyl substituted | | Plumbagin | 411(3.76) |
| **13a** | 347(3.15) | Trifluoromethanesulfonate-substituted | |
| **13b** | 338(2.82) | **15a** | 317(3.04) |
| **13c** | 348(2.45) | **15b** | 318(2.97) |
| **13d** | 335(2.76) | Diethylphosphate substituted | |
| **13e** | 345(2.92) | **16a** | 326(2.71) |
| Methoxy substituted | | **16b** | 327(2.78) |
| **10a** | 330(2.06)/*383(1.86)* | Miscellaneous | |
| **10b** | 321(2.22)/*382(1.70)* | **17** | 334(2.54) |
| **10c** | 351(2.57)/*424(1.70)* | **18** | 330(2.62) |
| Halogen substituted | | | |
| **11a** | 314(0.92) | | |
| **11b** | 327(2.63) | | |
| **10d** | 319(2.15) | | |

[a] DMSO, $T = 25.0(2)^\circ$C; The errors on the $\lambda_{max}$ and on the $\varepsilon$ have been estimated to ±0.5 nm and 10%, respectively. The values in italics correspond to the ICT absorption.

The analysis of the electronic spectra of menadione and of its synthetic analogues revealed that substitution of the benzene moiety of 1,4-NQ significantly alters the energy of the quinoidal π–π* transition (HOMO/LUMO). The $\lambda_{max}$ measured at 331 nm for menadione is shifted from +20 nm for **10c** (6,7-dimethoxy) to −17 nm for **11a** (6-fluoro). For a same substituent ($CH_3$ or $OCH_3$, Table 1), the position of its substitution also modulates the 1,4-NQ π–π* energy. Due to its electron-donating inductive character (+I), the methyl groups (the Supporting Information, Figure S2), substituted in position 6 or 7, increase the electron density of the 1,4-NQ chromophore and destabilize the HOMO orbital leading to a smaller HOMO/LUMO separation ($\Delta\lambda = +4$ and +7 nm for **13d** and **13b**, respectively) with respect to menadione. The latter is additive and a bathochromic shift of +16 nm is measured for the 6,7-dimethylated analogue. Strong steric effects of the 5- and 8-positions (**13c**, **13a**) are highlighted by the large bathochromic shift of the π–π* ($\Delta\lambda > +17$ nm/menadione).

These two positions are very sensitive to steric hindrance of any bulky substituent such as methyl groups (the Supporting Information, Table S3). Considering the 6- and 7-substituted series (the Supporting Information, Figure S1 and S3), clear impact of the electron-donating (+I, +M) and electron-withdrawing (−I) effects of the substituents on the 1,4-NQ π–π* transitions can be evidenced. The methoxylated derivatives (**10a** and **10b**) are characterized by the formation of an additional ICT absorption corresponding to a valuable probe of their mesomeric properties (+M). These electronic processes seemingly overwhelmed the inductive effect of the oxygen heteroatom (−I). The most important effects are observed for the trifluoromethanesulfonate analogues **15a**/**15b** (−M, −I, $\Delta\lambda \approx -14$ nm) and for the fluoro analogues **11a**/**10d** (−I, $\Delta\lambda \approx$ −13–17 nm) analogues. For the chloro- (**11b**), diethylphosphonate- (**16a**, **16b**) and acetoxy- (**18**) substitutions, weak hypsochromic shifts of the 1,4-NQ π–π* transition are measured with respect to menadione as a result of their moderate inductive electron-withdrawing effects. This spectroscopic approach then allowed us to assess the electronic effects of various substituents on the absorption properties of the 1,4-NQ chromophore. Lowering the electron density of menadione by inductive effects increases the HOMO/LUMO separation, whereas electron-donating groups (inductive or mesomeric) induces bathochromic shifts of the π–π* transitions.

### Electrochemical properties

In aprotic solvent, 1,4-NQ reduction occurs through two successive one-electron transfers. $^1$H NMR Diffusion-ordered spectroscopy (DOSY) and chronocoulometry/chronoamperometry (the Supporting Information, Figures S4 and S5) were used to confirm that each of successive reduction processes effectively involves one electron transfer. The radical anion intermediate 1,4-NQ$^{-\cdot}$ is then formed in the $E_{pc1}$ reduction step, and is subsequently reduced to quinone dianion 1,4-NQ$^{2-}$ in a second $E_{pc2}$ step. Table 2 gathers the voltammetric behavior of each of the 1,4-naphthoquinones studied in this work. Two consecutive one-electron quasi-reversible waves ($E^1_{pc} - E^1_{pa} \approx 68$–104 mV

**Table 2.** Electrochemical data measured using cyclic voltammetry (CV) for all the menadione derivatives examined in this work.[a]

| 1st Redox wave | | | 2nd Redox wave | | |
|---|---|---|---|---|---|
| Compound | $E_{pc}/E_{pa}$ | $E_{1/2}(\Delta E)$ | $E_{pc}/E_{pa}$ | $E_{1/2}(\Delta E)$ | $\Delta E_{1/2}$ |
| **M.** | −0.65/−0.56 | −0.61(93) | −1.35/−1.25 | −1.30(96) | 0.69 |
| **10a** | −0.68/−0.59 | −0.64(90) | −1.39/−1.31 | −1.35(82) | 0.71 |
| **10b** | −0.67/−0.58 | −0.63(88) | −1.41/−1.29 | −1.35(122) | 0.72 |
| **10c** | −0.69/−0.61 | −0.65(82) | −1.41/−1.32 | −1.36(88) | 0.71 |
| **10d** | −0.58/−0.49 | −0.54(92) | −1.28/−1.18 | −1.23(100) | 0.69 |
| **11a** | −0.57/−0.51 | −0.54(68) | −1.32/−1.14 | −1.23(180) | 0.69 |
| **11b** | −0.56/−0.46 | −0.51(92) | −1.26/−1.18 | −1.22(82) | 0.71 |
| **13b** | −0.68/−0.60 | −0.64(86) | −1.40/−1.31 | −1.35(87) | 0.71 |
| **13a** | −0.72/−0.63 | −0.68(86) | −1.44/−1.36 | −1.40(76) | 0.72 |
| **13c** | −0.71/−0.63 | −0.67(78) | −1.44/−1.35 | −1.40(92) | 0.73 |
| **13d** | −0.69/−0.58 | −0.63(104) | −1.41/−1.29 | −1.35(120) | 0.72 |
| **13e** | −0.70/−0.62 | −0.66(82) | −1.42/−1.33 | −1.38(92) | 0.72 |
| **P.** | −0.48/−0.39 | −0.44(90) | −1.10/−1.02 | −1.06(80) | 0.62 |
| **15a** | −0.49/−0.42 | −0.46(74) | −1.14/−1.05 | −1.10(90) | 0.64 |
| **15b** | −0.48/−0.40 | −0.44(80) | −1.13/−1.03 | −1.08(102) | 0.64 |
| **16a** | −0.60/−0.51 | −0.55(96) | −1.30/−1.22 | −1.26(83) | 0.71 |
| **16b** | −0.59/−0.51 | −0.55(86) | −1.30/−1.20 | −1.25(94) | 0.70 |
| **17** | −0.61/−0.53 | −0.57(75) | −1.29/−1.21 | −1.25(74) | 0.68 |
| **18** | −0.59/−0.52 | −0.55(77) | −1.27/−1.20 | −1.24(70) | 0.69 |

| | 3rd Redox wave | |
|---|---|---|
| | $E_{pc}/E_{pa}$ | $E_{1/2}$ |
| **15a** | −1.72/−1.54 | −1.63(185) |
| **15b** | −1.77/−1.47 | −1.62(292) |

[a] DMSO; $I$ = 0.1 M $n$-$Bu_4NPF_6$. ($E_{pc}$ and $E_{pa}$ (V), $E_{1/2}=(E_{pc}+E_{pa})/2$ (V), $\Delta E = E_{pa}-E_{pc}$ (mV), $\Delta E_{1/2}=E^1_{1/2}-E^2_{1/2}$ (V)). **M.** = menadione; **P.** = Plumbagin. $v$ = 200 mV $s^{-1}$; reference electrode = KCl(3 M)/Ag/AgCl; working electrode = glassy carbon disk of 0.07 $cm^2$ area; auxiliary electrode = Pt wire.

and $E^2_{pc}-E^2_{pa}\approx$70–180 mV) have been observed for all the studied compounds.

In addition to these two quasi-reversible steps, a third redox wave (broad and ill-defined cathodic and anodic peaks) at more negative potential ($E^3_{1/2}\approx-1.6$ V) can be observed for the triflate-substituted menadiones (**15a** and **15b**), which likely results from the reduction of the aryl dialkyl sulfonate moiety to afford the radical anion of the sulfonic ester[71] (Figure 3).

The presence of a 5-hydroxyl group for plumbagin does not affect the reversibility and the shape of the two one-electron reduction processes, which indicates that these ionizable substituents are not acidic enough to protonate the two reduced species. In addition, the intramolecular hydrogen bonding[72–75] stabilizes the negative charge of the reduced species and renders these derivatives more oxidant than menadione.

The current intensities of the second reduction step were found to be weaker with respect to that leading to the radical anion intermediate 1,4-$NQ^{\bullet-}$ (the Supporting Information, Table S4). This electrochemical behavior might result from either comproportionation reactions ($E^1_{1/2}-E^2_{1/2}\approx$0.62–0.73 V and ln $K_{Comp}$ are ranging from 24 to 29)[76–79] or by fast and irreversible dimerization process[80] between the quinone dianion (1,4-$NQ^{2-}$) and the naphthoquinone (1,4-NQ) to afford an electro-inactive dimeric species[84] 1,4-$NQ_2^{2-}$.

With some exceptions (**15a** and **15b**), the degree of reversibility for the first electron transfer is indicated by the ratio $i_{pa}/i_{pc}$, which is close to 1 at $v$ = 200 mV $s^{-1}$, whereas those of the second redox process are often close to 0.8 (the Supporting Information, Table S4). The cathodic $i_{pc}$ and anodic $i_{pa}$ intensities also vary linearly[81] with the square root of the potential scan rate $v^{1/2}$: $i_p=(2.69\times10^5)\,n^{3/2}\,A\,D^{1/2}$ [1,4-NQ] $v^{1/2}$, which confirms that these electron transfers are diffusion-controlled ($D\approx0.38-1.55\times10^{-6}$ $cm^2\,s^{-1}$, the Supporting Information, Table S5).[82] These values are in quite good agreement with those determined for **10a** in [$D_6$]DMSO by $^1$H NMR DOSY (the Supporting Information, Figure S4). The voltage scan rate has, however, no influence on the reduction and oxidation potentials. Except for **15a** and **15b**, the general diffusion sequence ($pa_1 > pc_2 > pc_1 > pa_2$) suggests that the radical anion intermediate migrates by diffusion within the electrode better than the quinone dianion. This might explain in part the fact that weaker intensities are measured for the second redox step and can be ascribed to the negative charges borne by the two reduced species 1,4-$NQ^{\bullet-}$ and 1,4-$NQ^{2-}$.

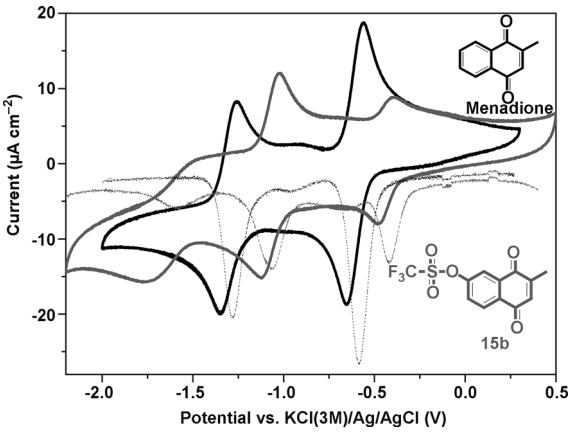



**Figure 3.** CV and SWV profiles of menadione (1.00 mM) and its 7-triflate substituted analogue **15b** (1.01 mM) measured in DMSO with 0.1 M $n$-$Bu_4PF_6$ electrolyte support at 25 °C. $v$ = 200 mV $s^{-1}$; reference electrode = KCl (3 M)/Ag/AgCl; working electrode = glassy carbon disk of 0.07 $cm^2$ area; auxiliary electrode = Pt wire.

We also evaluated the electrochemical properties of our menadione analogues using square wave voltammetry (SWV, the Supporting Information, Figures S6–S8). SWV is a rapid, sensitive, simple, and accurate analytical technique and allowed to reject background currents. Importantly, it allows a direct and accurate measurement of $E_{1/2}$ for (quasi)reversible reactions even with proximate potential values (Figure 3). Table S6 (the Supporting Information) summarizes the main results obtained from this approach. The electrochemical data obtained by these two methods (CV and SWV) were found to be in excellent agreement.

The redox potentials measured for the methylated series (Table 2) were compared to data available in the literature.[83] To the best of our knowledge, these values (only $E^1_{1/2}$ values were considered) are the only ones measured for menadione derivatives substituted on the benzene moiety. Despite the different experimental conditions (50 mM $n$-$Bu_4BF_4$ in DMF, $v$ = 100 mV $s^{-1}$), an excellent agreement was found (the Supporting Information, Table S7). These results therefore indicate that

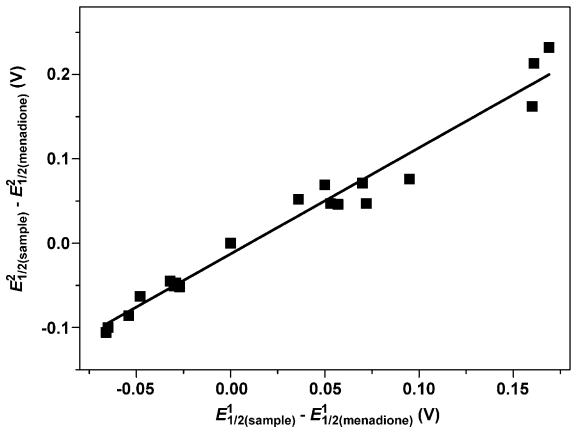



**Figure 4.** Variation of $\Delta E^2_{1/2}$ (V, second redox step) as a function of $\Delta E^1_{1/2}$ (V, first redox step). $\Delta E^n_{1/2} = E^n_{1/2(sample)} - E^n_{1/2(menadione)}$ with $n = 1,2$ ($\Delta E^2_{1/2} = -0.0127 + (1.259 \times \Delta E^2_{1/2})$, $R^2 = 0.97$). $v = 200$ mV s$^{-1}$; reference electrode = KCl(3 M)/Ag/AgCl; working electrode = glassy carbon disk of 0.07 cm$^2$ area; auxiliary electrode = Pt wire.

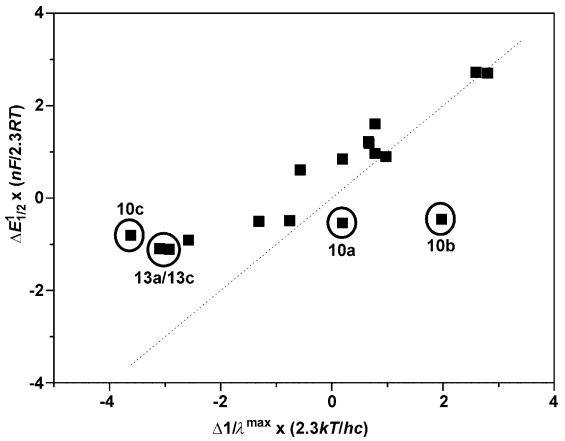



**Figure 5.** Variation of $\Delta E^1_{1/2}$ ($\times nF$/2.3RT) as a function of $\Delta(1/\lambda_{max}) \times (2.3kT/hc)$. $\Delta E^1_{1/2} = E^1_{1/2(sample)} - E^1_{1/2(menadione)}$ and $\Delta 1/\lambda^{max} = 1/\lambda_{max}^{menadione} - 1/\lambda_{max}^{sample}$. Plumbagin have been excluded from this data analysis. The dotted line is only a guide for the eyes and represents the expected relationship between $E^1_{1/2}$ and $1/\lambda_{max}$. $v = 200$ mV s$^{-1}$; reference electrode = KCl(3 M)/Ag/AgCl; working electrode = glassy carbon disk of 0.07 cm$^2$ area; auxiliary electrode = Pt wire.

DMF and DMSO (solvents with comparable properties) and the tetrabutylammonium salts ($PF_6^-$ and $BF_4^-$) might be assumed as inert partners during the electron transfers.

A clear relationship (Figure 4) can be observed between the half-wave potentials ($E^1_{1/2(sample)} - E^1_{1/2(menadione)}$ and $E^2_{1/2(sample)} - E^2_{1/2(menadione)}$) of the first and second electrochemical process. This strongly suggests that the substituents borne by the menadione core have the same electronic effects both on the 1,4-NQ unit or on its one-electron-reduced 1,4-NQ$^{\bullet-}$. Therefore, we anticipated that substituents effects would affect the same way the electrochemical properties of both the radical anion intermediate 1,4-NQ$^{\bullet-}$and the quinone dianion 1,4-NQ$^{2-}$.

By using the Hammett free energy relationships [Eq. (2) and (3)], it might also be possible to correlate the difference in the first half-wave potentials ($E^1_{1/2}$) with the difference of $\lambda_{max}$ [54,74]

$$\Delta E^1_{1/2} = E^1_{1/2(sample)} - E^1_{1/2(menadione)} = \frac{2.3RT}{nF}\rho_R \Sigma\sigma_{m/p} \quad (2)$$

$$\Delta\frac{1}{\lambda_{max}} = \frac{1}{\lambda_{max}^{sample}} - \frac{1}{\lambda_{max}^{menadione}} = \frac{hc}{2.3kT}\rho_R \Sigma\sigma_{m/p} \quad (3)$$

in which $\sigma$ are the Hammett substituent constants ($m$ = *meta*-, $p$ = *para* position with respect to the reactive site) and $\rho_R$ is the reaction constant.[34,35,84] If the reactivity of the substituted menadiones toward the first electron transfer can be probed by both their absorption and electrochemical properties, the plot of $[\Delta E^1_{1/2}(\times nF/2.3RT)]$ as a function of $[\Delta 1/\lambda_{max} \times (2.3kT/hc)]$ should provide a linear relationship with zero intercept (symbolized by the dot line in Figure 5). If we can consider that this hypothesis is obeyed for most of the compounds, we can, however, notice some relevant deviations. For example, plumbagin has not been considered in this data analysis. Besides, strong steric effects induced by methyl substitution of the 5- and 8-positions (**13c**, **a**) markedly alters the electronic properties of the 1,4-NQ core and therefore displays peculiar absorption and electrochemical behaviors. The methoxylated compounds (**10a**, **b**, and **c**) are characterized by mesomeric effects (ICT absorption), which deeply modify their absorption properties. This resonance effect also confers to the methoxylated derivatives specific electrochemical properties.

### Substitution effect on the menadione electrochemical properties

Similarly to the absorption study, we evaluated the electronic effects of the substituents on the electrochemical properties of the redox-active 1,4-NQ skeleton. The electrochemical behavior of the menadione derivatives in DMSO involves two distinct reduction steps leading to the restrained stable intermediates, the radical anion intermediate, 1,4-NQ$^{\bullet-}$, and the quinone dianion, 1,4-NQ$^{2-}$. The reactivity of these intermediate species depends basically on the quinone electronic properties, which alternatively can be microscopically described by the HOMO–LUMO energy or, empirically, by the Hammett[35,51] $\sigma$ constant (Figure 5). The menadione reactivity can be modified by adding electron-releasing or -withdrawing substituents to the 1,4-NQ electrophore. The half-wave potential of reference menadione in our study ($E^{1/2}_{menadione}$) is shifted to a new value ($E^{1/2}_{sample}$) by introducing one or several substituents (Figure 6). When steric interactions are excluded (5- and 8-methylated menadiones for instance, Figure 6), this shift is expected to be influenced solely by the electronic effects of the particular substituents and can be theoretically predicted by the Hammett–Zuman relationship [Eq. 3]. In this Equation, $\sigma_X$ corresponds to the electronic substituent constant of a substituent X ($\sigma_m$ stands for a *meta*-substituted group with respect to the reactive center and $\sigma_p$ designates a *para*-substituted function), whereas the reaction constant $\rho_R$ reflects the susceptibility of the reaction toward electronic perturbation brought by substitution. The vitamin K3 analogues are peculiar systems since the substituent located on the benzene ring (position 6 or 7)

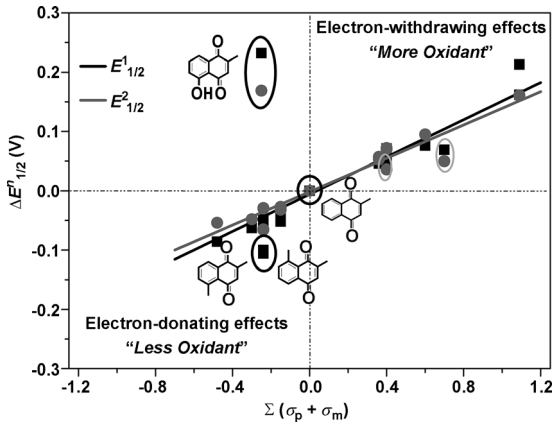



**Figure 6.** Plots of $\Delta E^1_{1/2}$ (■) and $\Delta E^2_{1/2}$ (●) versus $\Sigma(\sigma_m+\sigma_p)$ for the substituted menadione series. Solvent: DMSO; $I$ = 0.1 M $n$-$Bu_4PF_6$; $T$ = 25 °C; $v$ = 200 mV $s^{-1}$; reference electrode = KCl(3 M)/Ag/AgCl; working electrode = glassy carbon disk of 0.07 $cm^2$ area; auxiliary electrode = Pt wire. The light-gray circles (○) designate the compounds for which the Hammett constants have been estimated. The black circles (○) designate the compounds for which deviation from the linear relationship occurs.
$\Delta E^1_{1/2} = 0.130(1)\times(\Sigma(\sigma_m+\sigma_p)) - 0.006(4)$ ($R^2 = 0.954$);
$\Delta E^2_{1/2} = 0.141(8)\times(\Sigma(\sigma_m+\sigma_p)) - 0.002(4)$ ($R^2 = 0.952$)

can be considered as being either a *meta* substituent with respect to one carbonyl unit or a *para* substituent with respect to the other carbonyl one. As the quinone moiety is fully conjugated, any electronic effect can be, in principle, spread to each of the two carbonyls. Consequently, we used in our data analysis the sum $(\sigma_m+\sigma_p)$ to describe the electronic effect of a substituent on the electroactive menadione core (Figure 6). In the case of polysubstituted menadiones (**13e** and **10c**), the contribution of each substituent were obviously considered. Figure 6 illustrates the variation of $\Delta E^n_{1/2}$ ($n$ = 1 or 2) for the first and second electron transfers. Except for some menadione derivatives (plumbagin, **13c**, and **13a**), a clear linear variation of $\Delta E^1_{1/2}$ and $\Delta E^2_{1/2}$ can be observed with the electronic substituent constants ($\Sigma(\sigma_m+\sigma_p)$). In this data processing, the unavailable $\sigma_m$ and $\sigma_p$ parameters (–OP(O)$(OC_2H_5)_2$, **15a** and **15b**; –OC(O)N$(CH_3)_2$, **17**) were estimated by calculating the p$K_a$ values of *meta*- and *para*-substituted benzoic acids by using a ChemAxon p$K_a$ calculator.[85] The observed variations suggest that the electroactive menadione core is solely influenced by inductive or mesomeric electronic effects. Electron-withdrawing substituents decrease the electronic density of the 1,4-NQ unit and render these compounds "more oxidant" (easier to be reduced), whereas the electron-donating substituents increase the electronic density and thereby decrease the propensity of the corresponding redox compounds to be reduced. Interestingly, these electronic and electrochemical properties are similarly shared by both the menadione parent compound as well as its one-electron-reduced species as depicted by the comparable reaction constant $\rho_R$ ($\rho^1_R = 2.20$ and $\rho^2_R = 2.66$, 1 and 2 stands for the first and second reduction processes, respectively[86]). In the strict implementation of the Hammett–Zuman Equation, its intercept should be null. This is the case, within experimental errors, for $\Delta E^1_{1/2}$ and $\Delta E^2_{1/2}$ (Figure 6). This fact hints that the two consecutive electron transfers are not complicated by coupled chemical reactions such as proton transfer or complexation. The positive value of $\rho_R$ implies that the reaction is facilitated upon lowering the electron density of the electrophore, and thus demonstrates that the electron-accepting capacity of these menadione derivatives linearly vary with given substituents electronic perturbation. Deviation from such a linear relationship, however, occurs when steric interactions (between the methyl substituent and the carbonyl function of the quinone) influence the quinone reduction such as observed for **13c** and **13a**. This results in loss of coplanarity and, as a consequence, in a decrease in conjugation and in more negative value of $E_{1/2}$. Deviation is also observed for plumbagin for which an intramolecular hydrogen bond stabilizes the negative charge of the reduced products and makes these derivatives more oxidant-like.

**Table 3.** Training and cross-validation parameters of global models based on the three considered descriptor spaces.[a]

| Descriptors set | RMSE | $R^2$ | $RMSE_{XV}$ | $Q^2$ |
|---|---|---|---|---|
| CODESSA descriptors | 0.043 | 0.96 | 0.065 | 0.91 |
| EED | 0.054 | 0.94 | 0.086 | 0.85 |
| ISIDA fragment counts | 0.046 | 0.96 | 0.108 | 0.77 |

[a] See the Experimental Section in the Supporting Information for details on the descriptors and the meaning of the reported parameters: RMSE = Root Mean Squared Error between calculated and measured redox potentials, in Volts; $R^2$ and $Q^2$ = training and cross-validated correlation coefficients.

### Modeling studies: Cross-validated models

Table 3 reports the statistical parameters of the global consensus models obtained with the three different descriptor sets. Fitted $R^2$ values approach 1.0, showing that it is very easy to find a linear approximation of the redox potential as a function of either descriptors. Most relevant, it is possible (in all descriptor spaces) to achieve this with all molecules simultaneously, quinones and indolone *N*-oxides combined. Encoding reactivity-related aspects with EEDs is not bound to a given chemotype. A linear law explaining 94% of the variance of the redox potential in terms of EED does exist within this composite set of 67 molecules, the RMS error being of only 0.05 V. The models are also seen to be quite robust in terms of the extensive cross-validation challenge, as the increase of $RMSE_{XV}$ with respect to training RMSE values never exceeds 50% of the latter.

Whereas EED terms require a key atom recognition and labeling step to pinpoint the redox centers of considered molecules, CODESSA and ISIDA terms are not explicitly focusing on the reactive centers. CODESSA provides global molecular descriptors including, some QC terms of putatively high relevance for redox potential modeling. ISIDA fragment counts provide the most empirical approach. It relies on a machine-learning based decomposition of redox potential variations into empirical increments associated to present fragments. Ac-

cordingly, the fragment count-driven models are the least stable in cross-validation. However, they have a triple advantage:

1) Simplicity and rapidity of descriptor calculations–no extra effort to label reactive centers is needed.
2) Availability of a simple applicability domain[87–89] checking scheme: fragment control. Accordingly, prediction of compounds missing key fragments (which are not quinones or indolones, for example) or molecules with fragments never encountered during the training phase may be rejected.
3) Our web server supports easy deployment of ISIDA models, which were thus made available on http://infochim.u-strasbg.fr/cgi-bin/predictor.cgi. It supports MarvinSketch input of a structure, or submission of a multi-molecule file. In return, the user receives a spreadsheet associating, for each molecule, the redox potential value predicted by each partial Equation contributing to the global consensus model. For each of these, a status column indicates whether the current compound counts as within or without the applicability domain of the given partial equation. Eventually, average and standard deviations are returned both for 1) all predictions by partial equations and 2) only for those partial equations allegedly containing the molecule within their applicability domain.

**Table 4.** Predictive challenge (four newly synthesized molecules). ISIDA (AD) = ISIDA models with applicability domain checking (fragment control).

| | | | $E^1_{1/2}$ [V][a] | | |
|---|---|---|---|---|---|
| Compound | Exptl | EED | CODESSA | ISIDA | ISIDA (AD) |
| **10 d** | −0.514 | −0.506(32) | −0.543(28) | −0.590(7) | −0.500(8) |
| **11 a** | −0.512 | −0.506(32) | −0.531(32) | −0.590(7) | −0.500(8) |
| **17** | −0.548 | −0.518(30) | −0.624(72) | −0.630(3) | −0.49(18) |
| **18** | −0.534 | −0.507(29) | −0.560(48) | −0.61(1) | −0.52(18) |

[a] The $E^1_{1/2}$ value (V/SCE) has been obtained from SWV measurements (the Supporting Information, Table S6) with the following parameters: $v$ = 20 mV s$^{-1}$, pulse height $\Delta E_p$ = 50 mV; pulse width $t_{step}$ = 50 ms; step potential $E_{step}$ = 5 mV, step amplitude = 1 mV. Reference electrode = KCl(3 M)/Ag/AgCl; working electrode = glassy carbon disk of 0.07 cm$^2$ area; auxiliary electrode = Pt wire.

### Prediction of redox potential of new menadione derivatives

The "PRED" molecules (Table 4) are a difficult test set, because of the subtle balance of herein involved electronic effects. In fluorine-substituted compounds, the electron-withdrawing inductive effect of −F is more or less counterbalanced by its donor resonance effect (note that the two analogues formally differ by a swapped −Me group position, so their similar redox potentials are no surprise). The acetate derivative features an oxygen bound to the benzoquinone, per se less electronegative, but also electron-depleted as part of an ester group, thus less involved in resonance. The measurement shows this molecule has slightly less affinity for electrons than the F analogues: The drop in inductive effect seems thus to have the upper hand. In the carbamate analogue, the loss of −O–Phe resonance effect strength should be no longer an issue, as the carbonyl group preferentially draws electrons from the conjugated −$NMe_2$: indeed, this compound is still less electron-affine.

Predictions shown in Table 4 report the average ± standard deviation of the redox potential (in V) as predicted by all the partial equations contributing to the consensus model (see the cross-validated model building in the Experimental Section in the Supporting Information). A large standard deviation means that the partial models, each calibrated on hand of randomized subsets of ALL according to the cross-validation procedure, return strongly divergent results when challenged to predict the redox potential of a new molecule. It is a reflection of the fact that calibrated partial models were strongly tributary on the peculiar training subsets, and thus the overall confidence of such predictions is intrinsically lower.

The EED model does, indeed, predict the observed trend in terms of relative electron affinity −F > −OAc > −OC(=O)N, and yet underestimates the relative redox potential shift. It considers the ester as roughly equally electron affine as the F− derivatives, and although correctly pointing out the trend induced by the carbamate substituent, it minimizes its actual impact. This notwithstanding, the EED model is never wrong by more than 0.03 V on an absolute scale, and, since this is its intrinsic imprecision as expressed by its standard deviation, prediction by the EED model should count as correct.

CODESSA-driven equations are also successful in the above-mentioned sense. They reproduce the ranking in terms of relative redox potentials correctly, yet by contrast to EED they overemphasize the relative impact of substituents. Nevertheless, predicted values are roughly within one standard deviation from experimental measures, except that these standard deviations are significantly larger than the ones of EED models.

Interestingly, the ISIDA approach behaves, in terms of predicted relative trends, very much like the EED approach: Correct ranking, but predicting a lower-than-observed redox potential shift between the most and less electron-affine chemical entity. However, the predicted potentials are all systematically shifted towards more negative values. Furthermore, if applicability domain filters are activated (last column) only the partial equations that had access to molecules "similar" to the predicted one are taken into account. This should, normally, improve the results, knowing that "similar" in the above sense practically means "sharing the fragments associated to the descriptors entering the model". In practice, only equations trained on subsets containing at least one fluorinated quinone would serve to predict the first two fluoro derivatives, and so on. In practice, the number of models qualifying for prediction under the applicability domain constraint turned out to be very low; hence, large standard deviations and less accurate averages.

## Conclusion

The herein-developed QSPR models represent a first successful modeling attempt of the redox potential applicable to several different redox-active scaffolds. They succeeded in cross-valida-

tion, and managed to correctly rank newly synthesized analogues in terms of their relative electron affinities. These predictions were quantitatively accurate, experimental and predicted values being within the nominal accuracy range of the models. However, extremely accurate predictions of very subtle redox potential shifts of the order of a few millivolts remain beyond the ability of simple QSPR approaches. The ISIDA fragment counts-based model (not the most accurate, but the fastest and easiest to use) was deployed on our QSPR web server, and is freely accessible.

Predictive models may well be successful even though they fail to highlight the underlying mechanisms of a process. This is because they rely on observed correlations, and correlations do not imply causal relationships.[56] Hence, models using a set of descriptors tailor-made to capture the through-bond effects of the rest of the molecule on the electron density of atoms allegedly involved in the electron-transfer process (the carbonyl carbons) turned out to successfully explain the so-far available experimental redox potentials within two distinct series (quinones and indolone *N*-oxides).

The herein on-purpose developed Electronic Effect Descriptors (EED) allowed an explanation of the redox potentials of both families within a common model. Extending the scope of modeling beyond the quinone chemotype, as in preceding works, was actually proven to benefit machine learning. Without information from the indolone *N*-oxide set, none of the approaches would have been able to "learn" redox potential modeling sufficiently well to be able to properly predict the herein reported new menadione analogues, except for plumbagin. The latter failure can be well explained, being the only case featuring an intramolecular hydrogen bond. The impact thereof on the redox potential cannot be properly "learned" in absence of similar training examples. Such an effect should be, in principle, properly handled by QC descriptors derived from basis state geometries. Unfortunately, such terms have their own intrinsic limitations, representing a trade-off between the accuracy level of the quantum calculation and the allotted computer time.

## *Acknowledgements*

The authors wish to thank the International Center for Frontier Research in Chemistry (ic-FRC) in Strasbourg (ic-FRC-LabEx Chimie des systèmes complexes, project entitled "Understanding the mechanisms of antimalarial redox-active substrates in Plasmodium-infected red blood cells: a combined physicochemical and computational approach to unveiling biological complexity") for creating a proper framework for this scientific collaboration. The Centre National de la Recherche Scientifique (CNRS) and the University of Strasbourg (UMR 7509 CNRS-UdS), the ANRémergence program (grant SCHISMAL, E.D.C.), the Laboratoire d'Excellence (LabEx) ParaFrap (grant LabEx ParaFrap ANR-11-LABX-0024, E.D.C.), partly supported this work. The authors (M.E., E.C.R., D.A.L., and E.D.C) are grateful to Laetitia Robuchon-Tardif, Caroline Bayart and Laura Chamand for their help in measuring absorption and electrochemical data. P.S. and A.V. thank the Program of Competitive Growth of Kazan Federal University for support.

**Keywords:** chemoinformatics · cyclic voltammetry · electrochemistry · redox chemistry · structure–property relationships

[1] E. A. Hillard, F. C. de Abreu, D. C. M. Ferreira, G. Jaouen, M. O. F. Goulart, C. Amatore, *Chem. Commun.* **2008**, *0*, 2612.
[2] P. Kovacic, R. Somanathan, *Birth Defects Res. Part C* **2006**, *78*, 308.
[3] J. L. Bolton, M. Trush, T. Penning, G. Dryhurst, T. J. Monks, *Chem. Res. Toxicol.* **2000**, *13*, 135.
[4] T. J. Monks, D. C. Jones, *Curr. Drug Metab.* **2002**, *3*, 425.
[5] C. Morin, T. Besset, J.-C. Moutet, M. Fayolle, M. Bruckner, D. Limosin, K. Becker, E. Davioud-Charvet, *Org. Biomol. Chem.* **2008**, *6*, 2731.
[6] C. Biot, H. Bauer, R. H. Schirmer, E. Davioud-Charvet, *J. Med. Chem.* **2004**, *47*, 5972.
[7] K. Blumenstiel, R. Schöneck, V. Yardley, S. L. Croft, R. L. Krauth-Siegel, *Biochem. Pharmacol.* **1999**, *58*, 1791.
[8] L. Salmon-Chemin, E. Buisine, V. Yardley, S. Kohler, M.-A. Debreu, V. Landry, C. Sergheraert, S. L. Croft, R. L. Krauth-Siegel, E. Davioud-Charvet, *J. Med. Chem.* **2001**, *44*, 548.
[9] N. K. Cenas, D. Arscott, C. H. Williams, Jr., J. S. Blanchard, *Biochemistry* **1994**, *33*, 2509.
[10] N. Cenas, H. Nivinskas, Z. Anusevicius, J. Sarlauskas, F. Lederer, E. S. J. Arnér, *J. Biol. Chem.* **2003**, *279*, 2583.
[11] A. N. Kuntz, E. Davioud-Charvet, A. A. Sayed, L. L. Califf, J. Dessolin, E. S. J. Arnér, D. L. Williams, *PLoS Med.* **2007**, *4*, e206.
[12] M. Vos, G. Esposito, J. N. Edirisinghe, S. Vilain, D. M. Haddad, J. R. Slabbaert, S. Van Meensel, O. Schaap, B. De Strooper, R. Meganathan, V. A. Morais, P. Verstreken, *Science* **2012**, *336*, 1306.
[13] O. Kayser, A. F. Kiderlen, H. Laatsch, S. L. Croft, *Acta Trop.* **2000**, *76*, 131.
[14] L. Salmon-Chemin, A. Lemaire, S. De Freitas, B. Deprez, C. Sergheraert, E. Davioud-Charvet, *Bioorg. Med. Chem. Lett.* **2000**, *10*, 631.
[15] T. Müller, L. Johann, B. Jannack, M. Bruckner, D. A. Lanfranchi, H. Bauer, C. Sanchez, V. Yardley, C. Deregnaucourt, J. Schrevel, M. Lanzer, R. H. Schirmer, E. Davioud-Charvet, *J. Am. Chem. Soc.* **2011**, *133*, 11557.
[16] W. Friebolin, B. Jannack, N. Wenzel, J. Furrer, T. Oeser, C. P. Sanchez, M. Lanzer, V. Yardley, K. Becker, E. Davioud-Charvet, *J. Med. Chem.* **2008**, *51*, 1260.
[17] E. Davioud-Charvet, S. Delarue, C. Biot, B. Schwöbel, C. C. Boehme, A. Mössigbrodt, L. Maes, C. Sergheraert, P. Grellier, R. H. Schirmer, K. Becker, *J. Med. Chem.* **2001**, *44*, 4268.
[18] P. Wardman, *Prediction and Measurement of Redox Properties of Drugs and Biomolecules in Selective Activation of Drugs by Redox Processes, Vol. 198* (Eds: G. E. Adams, A. Breccia, E. M. Fielden and P. Wardman), NATO ASI Series, Ferno, Italy, **1990**, pp. 11 – 24.
[19] H. Bauer, K. Fritz-Wolf, A. Winzer, S. Kühner, S. Little, V. Yardley, H. Vezin, B. Palfey, R. H. Schirmer, E. Davioud-Charvet, *J. Am. Chem. Soc.* **2006**, *128*, 10784.
[20] D. A. Lanfranchi, E. Cesar-Rodo, B. Bertrand, H.-H. Huang, L. Day, L. Johann, M. Elhabiri, K. Becker, D. L. Williams, E. Davioud-Charvet, *Org. BioMol. Chem.* **2012**, *10*, 6375.
[21] M. Veguillas, M. Ribagorda, M. C. Carreno, *Org. Lett.* **2011**, *13*, 656.
[22] M. C. Redondo, M. Veguillas, M. Ribagorda, M. C. Carreño, *Angew. Chem. Int. Ed.* **2009**, *48*, 370; *Angew. Chem.* **2009**, *121*, 376.
[23] C. Hansch, A. Leo, Exploring QSAR: Fundamentals and Applications in Chemistry and Biology., *Am. Chem. Soc*. Washington D.C., **1995**.
[24] A. Varnek, D. Fourches, V. Solov′ev, O. Klimchuk, A. Ouadi, I. Billard, *Solvent Extr. Ion Exch.* **2007**, *25*, 433.
[25] A. Varnek, D. Fourches, F. Hoonakker, V. P. Solov′ev, *J. Comput.-Aided Mol. Des.* **2005**, *19*, 693.
[26] A. R. Katritzky, A. Lomaka, R. Petrukhin, R. Jain, M. Karelson, A. E. Visser, R. D. Rogers, *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 71.
[27] A. Tropsha, P. Gramatica, V. K. Gombar, *QSAR Comb. Sci.* **2003**, *22*, 69.
[28] F. Ruggiu, G. Marcou, A. Varnek, D. Horvath, *Mol. Inf.* **2010**, *29*, 855.
[29] A. Varnek, D. Fourches, D. Horvath, O. Klimchuk, C. Gaudin, P. Vayer, V. Solov′ev, F. Hoonakker, I. V. Tetko, G. Marcou, *Curr. Comput.-Aided Drug Des.* **2008**, *4*, 191.

[30] R. Todeschini, V. Consonni, R. Mannhold, H. Kubinyi, H. Timmerman, *Handbook of Molecular Descriptors*, (Eds: R. Manhold, H. Kubinyi, H. Timmerman), Wiley-VCH, Weinheim, Germany, **2008**.
[31] A. R. Katritzky, R. Jain, A. Lomaka, R. Petrukhin, M. Karelson, A. E. Visser, R. D. Rogers, *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 225.
[32] X. H. Ma, J. Jia, F. Zhu, Y. Xue, Z. R. Li, Y. Z. Chen, *Comb. Chem. High Throughput Screening* **2009**, *12*, 344.
[33] E. Alpaydin, *Introduction to Machine Learning*, MIT Press, Cambridge, Massachusetts, USA, **2004**.
[34] D. H. McDaniel, H. C. Brown, *J. Org. Chem.* **1958**, *23*, 420.
[35] L. P. Hammett, *J. Am. Chem. Soc.* **1937**, *59*, 96.
[36] C. Hansch, P. P. Maloney, T. Fujita, R. M. Muir, *Nature* **1962**, *194*, 178.
[37] K. Alizadeh, M. Shamsipur, *J. Mol. Struct.* **2008**, *862*, 39.
[38] J. Cape, M. Bowman, D. Kramer, *Phytochemistry* **2006**, *67*, 1781.
[39] M. Namazian, P. Norouzi, R. Ranjbar, *J. Mol. Struct.* **2003**, *625*, 235.
[40] M. Namazian, P. Norouzi, *J. Electroanal. Chem.* **2004**, *573*, 49.
[41] M. R. Hadjmohammadi, K. Kamel, P. Biparva, *J. Solution Chem.* **2011**, *40*, 224.
[42] P. Thanikaivelan, V. Subramanian, J. Raghava Rao, V. Unni Nair, *Chem. Phys. Lett.* **2000**, *323*, 59–70.
[43] A. Behesti, P. Norouzi, M. R. Ganjali, *Int. J. Electrochem. Sci.* **2012**, *7*, 4811.
[44] E. Cesar Rodo, K. Ehrhardt, M. Lanzer, D. L. Williams, E. Davioud-Charvet, D. A. Lanfranchi, unpublished results, **2014**.
[45] S. P. Kounaves, *Handbook of Instrumental Techniques for Analytical Chemistry*, Prentice Hall, Upper Saddle River, **1997**.
[46] M. D. Rozeboom, I. M. Tegmo-Larsson, K. N. Houk, *J. Org. Chem.* **1981**, *46*, 2338.
[47] P. Zuman, *Substituent Effects in Organic Polarography*, Plenum Press, New York, USA, **1967**.
[48] M. Aguilar-Martínez, G. Cuevas, M. Jimenez-Estrada, I. Gonzalez, B. Lotina-Hennsen, N. Macias-Ruvalcaba, *J. Org. Chem.* **1999**, *64*, 3684.
[49] E. D. Costa, M. T. Molina, F. C. de Abreu, F. D. D. Silva, C. D. Costa, W. Pinho, I. B. Valentim, B. Aguilera-Venegas, F. Perez-Cruz, E. Norambuena, C. Olea-Azar, M. O. F. Goulart, *Int. J. Electrochem. Sci.* **2012**, *7*, 6524.
[50] J. E. Heffner, C. T. Wigal, O. A. Moe, *Electroanalysis* **1997**, *9*, 629.
[51] C. Frontana, A. Vazquez-Mayagoitia, J. Garza, R. Vargas, I. Gonzalez, *J. Phys. Chem. A* **2006**, *110*, 9411.
[52] R. Schmid, F. Goebel, A. Warnecke, A. Labahn, *J. Chem. Soc. Perkin Trans. 2* **1999**, 1199.
[53] R. Svetlitski, A. Lomaka, M. Karelson, *Sep. Sci. Technol.* **2006**, *41*, 197.
[54] D. Horvath, F. Bonachera, V. S. Solov´ev, C. Gaudin, A. Varnek, *J. Chem. Inf. Model.* **2007**, *47*, 927.
[55] V. P. Solov´ev, A. Varnek, G. Wipff, *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 847.
[56] D. Horvath, *Rev. Roum. Chim.* **2010**, *55*, 783.
[57] M. P. Braban, I. Pop, X. Willard, D. Horvath, *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 1119.
[58] K. Reybier, T. H. Y. Nguyen, H. Ibrahim, P. Perio, A. Montrose, P. L. Fabre, F. Nepveu, *Biochemistry* **2012**, *88*, 57.
[59] J. K. Wegner, H. Frohlich, A. Zell, *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1077.
[60] C. Agostinelli, *J. Appl. Stat.* **2002**, *29*, 825.
[61] E. W. Steyerberg, M. J. Eijkemans, J. D. Habbema, *J. Clin. Epidemiol.* **1999**, *52*, 935.
[62] R. D. Clark, P. C. Fox, *J. Comput.-Aided Mol. Des.* **2004**, *18*, 563.
[63] D. M. Hawkins, S. C. Basak, D. Mills, *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 579.
[64] K. Baumann, *Trac-Trends in Analytical Chemistry* **2003**, *22*, 395.
[65] http://infochim.u-strasbg.fr/cgi-bin/predictor.cgi.
[66] K. A. Idriss, H. Sedaira, E. Y. Hashem, M. S. Saleh, S. A. Soliman, *Monatsh. Chem.* **1996**, *127*, 29.
[67] M. Umadevi, A. Ramasubbu, P. Vanelle, V. Ramakrishnan, *J. Raman Spectrosc.* **2003**, *34*, 112.
[68] E. A. Perpète, C. Lambert, V. Wathelet, J. Preat, D. Jacquemin, *Spectrochim. Acta Part A* **2007**, *68*, 1326.
[69] V. Prezhdo, O. Prezhdo, E. Ovsiankina, *Spectrochim. Acta Part A* **1995**, *51*, 2465.
[70] T. Mukherjee, *Proc. Ind. Natl. Sci. Acad., part A* **2000**, *66 (2)*, 239.
[71] J. C. Carnahan, W. D. Closson, J. R. Ganson, D. A. Juckett, K. S. Quaal, *J. Am. Chem. Soc.* **1976**, *98*, 2526.
[72] A. Kovács, A. Szabo, I. Hargittai, *Acc. Chem. Res.* **2002**, *35*, 887.
[73] L. S. Hernández-Muñoz, M. Gomez, F. J. Gonzalez, I. Gonzalez, C. Frontana, *Org. Biomol. Chem.* **2009**, *7*, 1896.
[74] N. Gupta, H. Linschitz, *J. Am. Chem. Soc.* **1997**, *119*, 6384.
[75] M. Aguilar-Martínez, N. A. Macias-Ruvalcaba, J. A. Bautista-Martinez, M. Gomez, F. J. Gonzalez, I. Gonzalez, *Curr. Org. Chem.* **2004**, *8*, 1721.
[76] E. N. da Silva Júnior, M. de Moura, A. V. Pinto, M. Pinto, M. de Souza, A. J. Araujo, C. Pessoa, L. V. Costa-Lotufo, R. C. Montenegro, M. O. de Moraes, V. F. Ferreira, M. O. F. Goulart, *J. Braz. Chem. Soc.* **2009**, *20*, 635.
[77] A. J. Araújo, A. A. de Souza, E. N. da Silva, J. D. B. Marinho, M. de Moura, D. D. Rocha, M. C. Vasconcellos, C. O. Costa, C. Pessoa, M. O. de Moraes, V. F. Ferreira, F. C. de Abreu, A. V. Pinto, R. C. Montenegro, L. V. Costa-Lotufo, M. O. F. Goulart, *Toxicol. in Vitro* **2012**, *26*, 585.
[78] P. S. Guin, S. Das, P. C. Mandal, *Int. J. Electrochem. Sci.* **2008**, *3*, 1016.
[79] D. O. Wipf, K. R. Wehmeyer, R. M. Wightman, *J. Org. Chem.* **1986**, *51*, 4760.
[80] M. W. Lehmann, D. H. Evans, *J. Electroanal. Chem.* **2001**, *500*, 12.
[81] J. E. B. Randles, *Trans. Faraday Soc.* **1948**, *44*, 327.
[82] J. Mauzeroll, A. J. Bard, *Proc. Natl. Acad. Sci. USA* **2004**, *101*, 14159.
[83] R. Schmid, F. Goebel, A. Warnecke, A. Labahn, *Proc. Natl. Acad. Sci. USA* **1999**, *96*, 1199.
[84] C. Hansch, A. Leo, R. W. Taft, *Chem. Rev.* **1991**, *91*, 165.
[85] ChemAxon, *pKa Calculator Plugin*, https://www.chemaxon.com/products/calculator-plugins/property-predictors/, **2007**.
[86] R. J. Driebergen, E. E. Moret, L. H. M. Janssen, J. S. Blauw, J. J. M. Holthuis, S. J. P. Kelder, W. Verboom, D. N. Reinhoudt, W. E. Vanderlinden, *Anal. Chim. Acta* **1992**, *257*, 257.
[87] D. Horváth, G. Marcou, A. Varnek, *J. Chem. Inf. Model.* **2009**, *49*, 1762.
[88] I. V. Tetko, I. Sushko, A. K. Pandey, H. Zhu, A. Tropsha, E. Papa, T. Oberg, R. Todeschini, D. Fourches, A. Varnek, *J. Chem. Inf. Model.* **2008**, *48*, 1733.
[89] R. P. Sheridan, B. P. Feuston, V. N. Maiorov, S. K. Kearsley, *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1912.

Received: June 2, 2014
Published online on December 30, 2014

ChemPubSoc Europe

DOI: 10.1002/cmdc.201600009

CHEMMEDCHEM Full Papers

SPECIAL ISSUE

# Redox Polypharmacology as an Emerging Strategy to Combat Malarial Parasites

Pavel Sidorov,[a, b] Israel Desta,[c, d] Matthieu Chessé,[c] Dragos Horvath,[a] Gilles Marcou,[a] Alexandre Varnek,[a] Elisabeth Davioud-Charvet,*[c] and Mourad Elhabiri*[c]

3-Benzylmenadiones are potent antimalarial agents that are thought to act through their 3-benzoylmenadione metabolites as redox cyclers of two essential targets: the NADPH-dependent glutathione reductases (GRs) of *Plasmodium*-parasitized erythrocytes and methemoglobin. Their physicochemical properties were characterized in a coupled assay using both targets and modeled with QSPR predictive tools built in house. The substitution pattern of the west/east aromatic parts that controls the oxidant character of the electrophore was highlighted and accurately predicted by QSPR models. The effects centered on the benz(o)yl chain, induced by drug bioactivation, markedly influenced the oxidant character of the reduced species through a large anodic shift of the redox potentials that correlated with the redox cycling of both targets in the coupled assay. Our approach demonstrates that the antimalarial activity of 3-benz(o)ylmenadiones results from a subtle interplay between bioactivation, fine-tuned redox properties, and interactions with crucial targets of *P. falciparum*. Plasmodione and its analogues give emphasis to redox polypharmacology, which constitutes an innovative approach to antimalarial therapy.

## Introduction

Drug combination for the treatment of infectious diseases such as malaria, AIDS, and tuberculosis is becoming a major strategy in the research and development of effective regimens, because drug resistance of pathogens spreads in many areas of the tropical world, making these diseases a major public health disaster. Effective combination regimens offer numerous benefits over single targeted monotherapy: 1) the risk of developing rapid resistance is high for drugs targeting a single step of a selected pathway, 2) the probability of inducing multiple individual target-based gene mutations is low for agents acting on multiple targets or whose targets are products of multiple genes, 3) the use of multitarget drugs or several drugs can achieve a therapeutic effect with greater efficacy than that achieved with monotherapy, and 4) the expected toxicity is decreased because of synergistic drug interactions that allow low effective doses. To illustrate these different advantages, one can refer to reported cases for recently marketed antimalarial drug combinations. Atovaquone, used alone, has rapidly (in a single year) induced resistance in *Plasmodium falciparum* parasites following mutations of cytochrome b in the region located at a putative drug binding site. Malarone (the combination of atovaquone and proguanil) rescued the marketed atovaquone development. Numerous artemisinin-based combination therapies (ACT) have become available and are widely used malaria drugs, owing to a significant drop (45%) in mortality and morbidity in malaria cases since 2000. These spectacular results are partly due to the massive use of oral ACT, designed to have a high cure rate and few side effects and to decrease malaria transmission. For example, coartem (the combination of artemether and lumefantrine), artekin (the combination of dihydroartemisinin and piperaquine), and artequin (the combination of artesunate and mefloquine), have been demonstrated in clinical practice to be very effective against drug resistance.

In addition to drug combinations, the identification of a multitarget drug is a promising and appealing strategy that may provide the desired drug discovery breakthrough required to treat malaria. The parasites need to develop resistance against one single agent that has numerous targets, exponentially delaying the onset of resistance, such as in a combination therapy. One example of the success of this research strategy is the recent discovery of pyrrolopyrazines as potent multitarget antimalarial agents[1] from in-depth bioinformatics analysis and target panel screening, suggesting IspD and multi-kinase inhibition. Therefore, it is believed that composite computational approaches may provide unique access to deciphering polypharmacological effects of new bioactive chemical agents.

[a] *P. Sidorov, Dr. D. Horvath, Dr. G. Marcou, Prof. A. Varnek*
*Laboratoire de Chemoinformatique*
*UMR 7140 CNRS–Université de Strasbourg*
*1 rue Blaise Pascal, Strasbourg 67000 (France)*

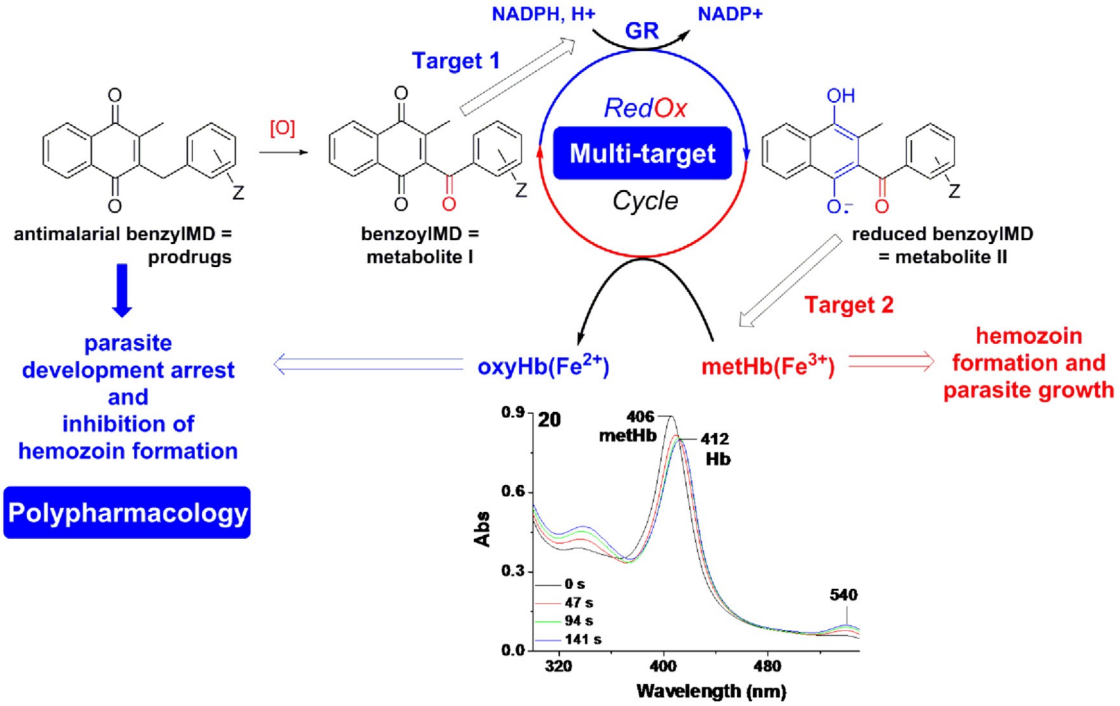
[b] *P. Sidorov*
*Butlerov Institute of Chemistry, Kazan Federal University*
*1/29 Lobachevskogo str., Kazan 420008 (Russia)*

[c] *I. Desta, M. Chessé, Dr. E. Davioud-Charvet, Dr. M. Elhabiri*
*Laboratoire de Chimie Bioorganique et Medicinale*
*UMR 7509 CNRS–Université de Strasbourg*
*European School of Chemistry, Polymers and Materials (ECPM)*
*25 Rue Becquerel, 67087 Strasbourg (France)*
*E-mail: elisabeth.davioud@unistra.fr*
*elhabiri@unistra.fr*

[d] *I. Desta*
*New York University Abu Dhabi (NYUAD), Saadiyat Island, Abu Dhabi (UAE)*

Supporting information and the ORCID identification number(s) for the author(s) of this article can be found under http://dx.doi.org/10.1002/cmdc.201600009.

SPECIAL ISSUE This article is part of a Special Issue on Polypharmacology and Multitarget Drugs. To view the complete issue, visit: http://onlinelibrary.wiley.com/doi/10.1002/cmdc.v11.12/issuetoc.




**Scheme 1.** Bioactivation of 3-benzylmenadiones (benzylMD) through a cascade of redox reactions, from their 3-benzoylmenadione (benzoylMD) metabolites, catalyzing both the glutathione reductase (Target 1) and methemoglobin (Target 2) redox cycling in *P. falciparum*-infected RBCs. The continuous conversion of metHb($Fe^{III}$) to oxyHb($Fe^{II}$) was spectrophotometrically evidenced by monitoring hemoglobin absorbance in a coupled assay using the NADPH-GR system in the presence of the redox cycler.

A unique class of multitarget agents is represented by redox-active agents: in their oxidized states, they can be reduced by Target 1 and regenerated by Target 2 catalysis through an oxidation step from their reduced state (Scheme 1). Methylene blue[2,3] and 3-[4-(trifluoromethyl)benzyl]menadione, henceforth called plasmodione,[3–5] are representatives of such redox cyclers, acting both as catalysts and NADPH-consuming turncoat inhibitors of multiple targets, that is, NADPH-dependent disulfide reductases[6] and methemoglobin($Fe^{III}$) (metHb)[2,7] with potent inhibition of *P. falciparum*-parasitized red blood cell (pRBC) growth in vitro,[3–5,8] moderate in vivo activities in orally administered *P. berghei*-infected mice,[4] and very low rates of spontaneous drug resistance.[8]

In their reduced state (i.e., the redox properties of plasmodione or methylene blue are seemingly the key parameters of their peculiar antimalarial activities), these redox cyclers play a critical role in multiple and interrelated vital processes of *P. falciparum* infecting RBCs. Above all, they can markedly alter the hemoglobin digestion process of the parasite by reversing the Hb($Fe^{II}$)↔metHb($Fe^{III}$) equilibrium toward the less digestible hemoglobin state. Alternatively, they can induce significant oxidative stress to the parasites by reacting with oxygen to lead to $O_2^{\bullet-}$/$H_2O_2$ species that will generate a hostile milieu in the presence of ferrous compound (see above, conditions for production of reactive oxygen species [ROS]). As these versatile compounds also act as efficient catalysts and effective substrates of NADPH-dependent disulfide reductases, they exhaust the *P. falciparum* antioxidant defense by overconsuming NADPH and generating a pro-oxidant environment deleterious to the parasite. Furthermore, we recently demonstrated[5] that these redox cyclers (or their metabolites), in oxidized or reduced states, continuously generate ROS in *P. falciparum* pRBCs, but can also alter hemoglobin catabolism. The generated ferrylhemoglobin($Fe^{IV}$) ultimately leads to hemichrome precipitation and subsequent and early phagocytosis of the pRBCs. Last but not the least, these redox cyclers can be converted into other metabolites of interest, acting on other processes vital to the parasite. Benzoxanthones, which have been described as efficient hemozoin inhibitors, can be produced in the pRBCs. Conferring fine-tuned redox properties to well-tailored compounds can represent a smart and efficient polypharmacological strategy to combat malarial parasites, although detailed mechanisms are not yet completely understood.

Therefore, building computational models to predict redox properties of putative antimalarial leads would undoubtedly speed up the development and delivery of multitarget redox-active drug candidates. The chemical space of possible analogues is virtually infinite, but resources for synthesis and property measurement are not. Synthesis can be assisted by computer-aided design of molecules with desired properties, based on routinely used quantitative structure-property relationship (QSPR) models,[9] to prioritize the molecules to be synthesized and tested. Such models are learned by experimental data;[10] structures with known desired properties are represented by molecular descriptors,[11–13] quantitative characteristics are calculated from the structure, then machine learning methods[14] are applied in order to establish a mathematical relationship between the experimental property (*P*) and descriptors (*D*).

Once this function is established, it may be applied on untested or virtual compounds to guide chemists toward molecules possessing desirable properties. The molecular descriptors may represent various characteristics of compounds, from physicochemical properties to particular structural motives and features.

Chemoinformatics-based prediction models of redox potential (especially concerning the key family of quinone compounds) are rare. Predictions of thermodynamic functions of the formation of the radical anion intermediate, as the first step of quinone reduction, using quantum chemical (QC) methods, have been more frequently reported. Often, the redox potential is predicted as a function of Gibbs free energy of oxidized and reduced forms, calculated by either Hartree–Fock[15–17] and DFT[18] calculations, or faster but less precise semiempirical methods. The influence of solvent on redox potentials of quinones was established as an MLR model,[19] based on QC descriptors and solvent properties (Guttmann acceptor number, dipole moment). Another QC characteristic, the electrophilicity index,[20] has been reported to robustly explain the redox potentials of a small dataset of quinones of different families (benzo-, naphtho-, and anthraquinones with various substituents).[21] However, the use of computationally heavy QC calculations limits the applicability of these models for virtual screening of a vast number of molecules. Therefore, a model based on simpler molecular descriptors is desirable. Recently, our group has reported the possibilities of QSPR modeling for prediction of the redox potential of substituted 2-methyl-1,4-naphthoquinones (menadiones).[22] An online evaluation model was built and made available to predict the electrochemical properties of these quinones and redox-active analogues to aid medicinal chemists in targeting their synthetic efforts toward redox-active agents.

In the present work, we extended our model to predict the oxidant character of polysubstituted 3-benzylmenadione derivatives and their putative metabolites for further optimization of the early lead plasmodione. First, all redox potential values of both sub-series of compounds were experimentally determined by cyclic and square-wave voltammetries. As 3-benzoylmenadiones act as catalytic inhibitors of both human and *P. falciparum* glutathione reductase (GR), we then applied a multitarget drug screening assay using the redox cycler in an in vitro reconstitution system in the presence of both targets, that is, Target 1: the human glutathione reductase (hGR), coupled to Target 2: metHb($Fe^{III}$). To biochemically probe the computationally predicted electron-transfer properties, we subjected polysubstituted 3-benzylmenadione derivatives to metHb reduction activity kinetics in the previously established coupled assay, based on the NADPH-dependent hGR and metHb.

## Results and Discussion

### Chemistry

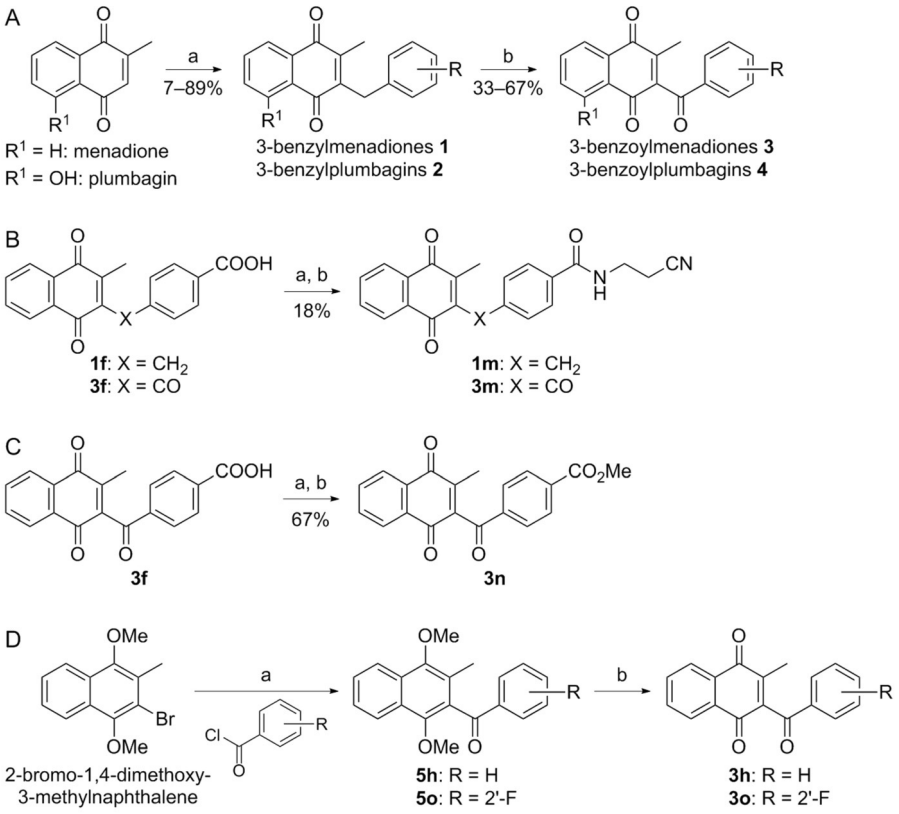
The silver-catalyzed decarboxylation reaction is an efficient method to synthesize 2-methyl-1,4-naphthoquinone derivatives starting from carboxylic acids, under previously described conditions.[23,24] In the present work, the silver-catalyzed coupling reaction represented an easy access to benzyl-substituted menadione derivatives **1**–**2**, including new representatives **1h**–**1u** (Scheme 2, route A). For the benzylic oxidation of 3-benzylmenadione, we used $CrO_3$ and periodic acid $H_5IO_6$ to obtain the desired 3-benzoylmenadiones **3**–**4** in moderate to good yields (33–67%); new representatives (**3h**–**3t**) are described here (Table S1). The reaction progress was easily followed by $^1$H NMR, as the singlet of the bridging $CH_2$ group disappeared over the course of the reaction until complete conversion.[4] To mask the polarity of the carboxy group of acids **1f** and **3f**, the corresponding amides were prepared (Scheme 2, route B).

Previous results indicated increased antimalarial activity for β-cyanoamides, prepared from acids and 3-aminopropionitrile, as intermediates in the synthesis of tetrazoles.[25] The acid was first converted into its acyl chloride with $SOCl_2$, and finally, 3-aminopropionitrile was added to afford amides **1m** and **3m**. Carboxylic acid **3f** was converted into its corresponding acid chloride and then stirred in MeOH to form methyl ester **3n** using a standard protocol (Scheme 2, route C). Finally, starting from the previously reported 2-bromo-1,4-dimethoxy-3-methylnaphthalene, accessible through bromination at the aromatic ring with $Br_2$ in $CH_2Cl_2$ at 0 °C (86%), the bromide was allowed to react with acyl chlorides (Scheme 2, route D). The resulting methanone intermediates **5h** and **5o** were oxidized by CAN to give the desired benzoyl derivatives **3h** and **3o**.

### Physicochemical investigations

We describe and discuss herein the physicochemical results that were obtained by first measuring the electrochemical properties (cyclic voltammetry [CV] and square-wave voltammetry [SWV]) of two homogenous series of 3-benzyl- and 3-benzoylmenadiones (Table 1, Table S2). These redox-active compounds differ by the substitution patterns of their benz(o)yl units (eastern region), which highlighted the key role of this moiety (electronic and steric effects) on the antimalarial activity of these derivatives.

Also, the benzoyl C=O group, which results from endogenous oxidation (bioactivation) of the benzylic related prodrugs, has a significant impact on redox properties. Complementarily to the electrochemical approach, we furthermore analyzed the capacity of the menadione analogues, in their reduced states, to efficiently reduce methemoglobin metHb($Fe^{III}$) to oxyhemoglobin oxyHb($Fe^{II}$) in a reduction assay coupled to the hGR/NADPH system under quasi-physiological conditions (Scheme 1). The latter continuously regenerated the reduced species of our substrates. This experiment reflected the capacity of the substrates to be efficiently reduced by hGR (Target 1) under NADP flux and to transfer its electron(s) to ferric species (Target 2). Together, these electrochemical, absorption spectroscopy, and kinetic data represent a valuable physicochemical dataset that highlights key aspects related to the mechanism of action of this class of multitarget compounds toward parasitic pathogens.




**Scheme 2.** Synthesis of benzyl- and benzoyl-substituted derivatives of menadione and plumbagin **1**–**4**. *Reagents and conditions*: A: a) phenylacetic acid derivative (2.0 equiv), $AgNO_3$ (0.1 equiv), $(NH_4)_2S_2O_8$ (1.3 equiv), $CH_3CN/H_2O$, 2 h, 85 °C; b) $CrO_3$ (0.2 equiv), $H_5IO_6$ (7.0 equiv), $CH_3CN$, 24 h, RT. B: a) $SOCl_2$ (8 mL), reflux, 2 h; b) 3-aminopropionitrile (1.0 equiv), $CH_2Cl_2$, RT, 1 h. C: a) $SOCl_2$ (10 equiv), reflux, 3 h; b) MeOH (5 mL), RT, 3 h. D: a) 1. *n*BuLi (1.1 equiv), dry THF, −78 °C, 15 min, 2. acyl chloride (1.2 equiv); b) CAN (3 equiv), $CH_3CN/H_2O$, 15 min, RT or 1. $BBr_3$ (1 equiv), $CH_2Cl_2$, 2. oxidation in open air.

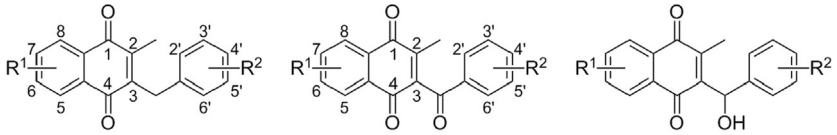
*Plasmodium* parasites use methemoglobin as a source of amino acids for their own growth and digest it more quickly than hemoglobin. Under the acidic conditions of the digestive food vacuole of *Plasmodium* parasites, hemoglobin ($HbFe^{II}$) or oxyhemoglobin (oxyHb or $oxyHb(Fe^{II})$) is quickly oxidized to methemoglobin (metHb or $metHb(Fe^{III})$). Consequently, the reduction of $metHb(Fe^{III})$ to $oxyHb(Fe^{II})$ can significantly slow metHb digestion. The ability of our redox cyclers to inhibit both human and plasmodial glutathione disulfide reductases (hGR, PfGR) of the parasitized erythrocytes, combined with their potency to reduce $metHb(Fe^{III})$ to $oxyHb(Fe^{II})$, contribute to a rise oxidative stress and interfere with hemozoin formation, with both phenomena leading to the death of the parasites. This pernicious and continuous futile NADPH flux under the catalysis of these disulfide reductases in the presence of redox-active compounds (which behave as substrates or electron acceptors) can dramatically increase the flux of toxic reduced species within the parasite and leads to its death by shifting the equilibrium $metHb(Fe^{III})$ to $oxyHb(Fe^{II})$ heme species.

## Electrochemical properties of benz(o)ylmenadiones

In this section, we describe the electrochemical properties of the 3-benz(o)ylmenadiones to evaluate the influence of the benzyl substitution. We previously evaluated the substitution effects on the menadione core (i.e., the western region of the 3-benz(o)ylmenadione derivatives) and found that the two one-electron transfers were sensitive to the nature of the substituent and the position of substitution.[22] The redox potentials of the substituted 3-benz(o)ylmenadiones (Table 1 and Figures S3–S42) were measured by CV (Table 2) and SWV (Table S3) at 25 °C using a glassy carbon electrode in DMSO solvent and tetra-*n*-butylammonium hexafluorophosphate $[N(nBu)_4PF_6]$ as the supporting electrolyte. These two electrochemical techniques are complementary, and the conditions are exactly similar to those employed for investigation on the substituted menadiones.[22] It is also noteworthy that the electrochemical properties measured by SWV (Table S3) were found to be in excellent agreement with those measured by CV. SWV is a rapid, sensitive, and accurate technique that enables subtraction of background currents. In addition, it allows direct and accurate measurement of $E_{1/2}$ for (quasi)reversible reactions, even with proximate potential values. The CV voltammograms were recorded across a potential range from +0.5 V to −2.2 V vs. KCl(3 M)/Ag/AgCl reference electrode (+0.210 V vs. NHE).[47]

In aprotic solvents, 1,4-naphthoquinone (1,4-NQ) reduction occurs through two successive one-electron transfers. The monoradical anion ($1,4\text{-}NQ^{\cdot-}$) is formed in the $E_{pc1}$ reduction step and is then reduced to its related dihydronaphthoquinone dianion ($1,4\text{-}NQ^{2-}$) in a second $E_{pc2}$ step (Figure 1). The 1,4-NQs

**Table 1.** Structures of 3-benz(o)ylmenadione (Benz(o)ylMD) derivatives, including 3-benzhydrolmenadione **6a**.




3-benzylMD **1**–**2** 3-benzoylMD **3**–**4** 3-benzhydrolMD **6**

| BenzylMD[a] | $R^1$ | $R^2$ | BenzoylMD[b] | $R^1$ | $R^2$ |
|---|---|---|---|---|---|
| **1a** | H | 4′-Br | **3a** | H | 4′-Br |
| **1b** | H | 4′-F | **3b** | H | 4′-F |
| **1c** | H | 4′-$CF_3$ | **3c** | H | 4′-$CF_3$ |
| **1d** | H | 4′-OH | **3d** | H | 4′-OH |
| **1e** | H | 4′-*t*Bu | **3e** | H | 4′-*t*Bu |
| **1f** | H | 4′-COOH | **3f** | H | 4′-COOH |
| **1g** | H | 4′-$NO_2$ | **3g** | H | 4′-$NO_2$ |
| **1h** | H | 4′-H | **3h** | H | 4′-H |
| **1i** | H | 4′-Me | **3i** | H | 4′-Me |
| **1j** | H | 4′-Cl | **3j** | H | 4′-Cl |
| **1k** | H | 4′-CN | **3k** | H | 4′-CN |
| **1l** | H | 4′-OMe | **3l** | H | 4′-OMe |
| **1m** | H | 4′-CONH$(CH_2)_2$CN | **3m** | H | 4′-CONH$(CH_2)_2$CN |
| | | | **3n** | H | 4′-$COOCH_3$ |
| | | | **3o** | H | 2′-F |
| **1p** | H | 2′-Br, 4′-$OCH_3$ | | | |
| **1q** | H | 2′-OMe | **3q** | H | 2′-OMe |
| **1r** | H | 3′,5′-$(OCH_3)_2$ | **3r** | H | 3′,5′-$(OCH_3)_2$ |
| **1s** | H | 2′,5′-$(OCH_3)_2$ | **3s** | H | 2′,5′-$(OCH_3)_2$ |
| | | | **3t** | H | 2′-F,5′-OMe |
| **1u** | H | 2′,3′,4′,5′,6′-$(F)_5$ | | | |
| **2a** | 5-OH | 4′-Br | **4a** | 5-OH | 4′-Br |
| | | | **6a**[c] | H | 4′-Br |

[a] Benzylmenadione scaffold: **1** and **2**. [b] Benzoylmenadione scaffold: **3** and **4**. [c] Benzhydrolmenadione scaffold: **6**.

**Table 2.** Electrochemical data (first and second redox waves for the benz(o)yl series; a third redox wave has been evidenced for the benzoyl series) measured by CV for all benz(o)ylmenadione derivatives examined in this work.[a]

| Compd | $E^1_{1/2}$(ΔE) [V(mV)] | $E^2_{1/2}$(ΔE) [V(mV)] | $\Delta E_{1/2}$ [V] | Compd | $E^1_{1/2}$(ΔE) [V(mV)] | $E^2_{1/2}$(ΔE) [V(mV)] | $E^3_{1/2}$(ΔE) [V(mV)] | $\Delta E_{1/2}$ [V] |
|---|---|---|---|---|---|---|---|---|
| menadione | −0.61(93) | −1.30(96) | 0.69 | | | | | |
| **1a** | −0.62(71) | −1.29(85) | 0.67 | **3a** | −0.46(78) | −1.12(80) | −1.63[b] | 0.66 |
| **1b** | −0.63(88) | −1.30(86) | 0.67 | **3b** | −0.46(86) | −1.14(82) | −1.65[b] | 0.68 |
| **1c** | −0.59(76) | −1.27(78) | 0.68 | **3c** | −0.43(80) | −1.12(84) | −1.62(60) | 0.69 |
| **1d** | −0.65(94) | −1.40(290) | 0.75 | **3d** | −0.53(80) | −1.17(80) | nd | 0.64 |
| **1e** | −0.64(80) | −1.38(80) | 0.74 | **3e** | −0.47(78) | −1.18(66) | −1.62[b] | 0.75 |
| **1f** | −0.65(80) | −1.35(109) | 0.7 | **3f** | −0.49(72) | −1.29(160) | −1.61[b] | 0.8 |
| **1g** | −0.60(70) | −1.38(80) | 0.74 | **3g** | −0.42(90) | −1.17(88) | −1.54[b] | 0.75 |
| **1h** | −0.63(80) | −1.33(128) | 0.70 | **3h** | −0.47(92) | −1.19(72) | −1.61[b] | 0.72 |
| **1i** | −0.64(86) | −1.35(93) | 0.71 | **3i** | −0.48(90) | −1.14(41) | −1.63[b] | 0.66 |
| **1j** | −0.62(84) | −1.30(84) | 0.68 | **3j** | −0.45(82) | −1.13(60) | −1.62[b] | 0.68 |
| | | | | **3k** | −0.43(80) | −1.09(70) | −1.39(86) | 0.66 |
| | | | | **3l** | −0.49(86) | −1.17(54) | nd | 0.68 |
| | | | | **3m** | −0.45(82) | −1.09(70) | br | 0.64 |
| | | | | **3n** | −0.44(88) | −1.10(92) | −1.74(156) | 0.66 |
| | | | | **3o** | −0.45(74) | −0.99(140) | −1.55[b] | 0.54 |
| **1q** | −0.65(80) | −1.36(94) | 0.71 | **3q** | −0.54(88) | −1.17(88) | br | 0.63 |
| **1r** | −0.65(82) | −1.35(82) | 0.7 | **3r** | −0.47(77) | −1.22(77) | −1.49[b] | 0.75 |
| | | | | **3s** | −0.54(96) | −1.16(88) | br | 0.62 |
| | | | | **3t** | −0.45(76) | −1.11(74) | −1.64(62) | 0.66 |
| **1u** | −0.60(80) | −1.25(132) | 0.65 | | | | | |
| plumbagin | −0.44(90) | −1.06(80) | 0.62 | | | | | |
| **2a** | −0.47(67) | −1.03(32) | 0.56 | **4a** | −0.30(82) | −0.93(76) | −1.49[b] | 0.63 |

[a] CVs measured in DMSO with $N(nBu)_4PF_6$ (0.1 M) electrolyte support at 25 °C; $v = 200$ mV $s^{-1}$; reference electrode: KCl(3 M)/Ag/AgCl; working electrode: glassy carbon disk with 0.07 $cm^2$ area. Compounds used for the validation test shown in blue. [b] SWV measurements; nd: not determined; br: broad signal.

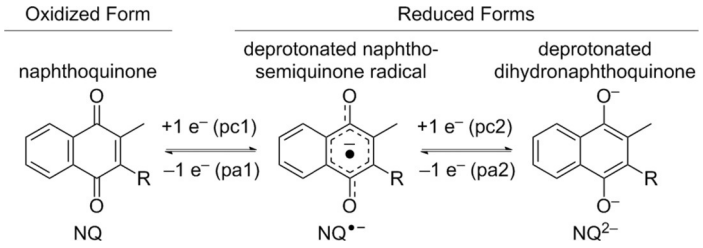



**Figure 1.** Oxidation/reduction processes of 1,4-NQs in aprotic solvents.

(except for the proton-donating systems, such as 3-benzylmenadiones **1 d** and **1 f**, and 3-benzoylmenadiones **3 d** and **3 f**, Tables 1 and 2) are not subjected to any inter- or intramolecular proton transfers and therefore the electrochemical properties are strictly associated with successive formation of the two anionic reduced species (Figure 1).

Substitution by either benzyl or benzoyl derivatives has no effect on the global electrochemical behavior of the 1,4-NQ core (i.e., two successive one-electron transfers). Table 2 shows the voltammetric data (CV) of each of the substituted 3-benz(o)ylmenadiones (including menadione[22] and plumbagin) investigated in this work. With the exception of hydroxylated menadiones, two consecutive one-electron quasi-reversible waves ($E^1_{pc}$-$E^1_{pa}$ ~67–96 mV and $E^2_{pc}$-$E^2_{pa}$ ~41–140 mV) were constantly observed. In addition to these two quasi-reversible steps for the benzylmenadiones series, a third redox wave at an intermediate potential value ($E^3_{1/2}$ ~ −1.07 V) was observed for 4′-nitrobenzylmenadione **1 g** that results from reduction of the nitro moiety (Figure 2).[26,27] Irrespective of the 3-benzoylmenadione derivative considered, a third (or fourth for **3 g**) weak, broad, and ill-defined wave was observed at more negative values ($E^3_{1/2} \gg -1.5$ V) that can be related to the oxidation/reduction of the benzoyl carbonyl unit (Table 2, Figure 3, and Table S3).

The degree of reversibility for the first electron transfer is indicated by the ratio $i_{pa}/i_{pc}$, which is close to 1 (at $v$ = 200 mV s$^{-1}$), while those in the second redox process are often close to 0.8. The cathodic $i_{pc}$ and anodic $i_{pa}$ intensities also vary linearly with the square root of the potential scan rate $v^{1/2}$ ($i_p = (2.69\times10^5)\ n^{3/2}\ A\ D^{1/2}$ [1,4-NQ] $v^{1/2}$),[28] which confirms that these electron transfers are diffusion-controlled processes.[29] The voltage scan rate, however, has no influence on the reduction and oxidation potentials, which further substantiates this property. The current intensities

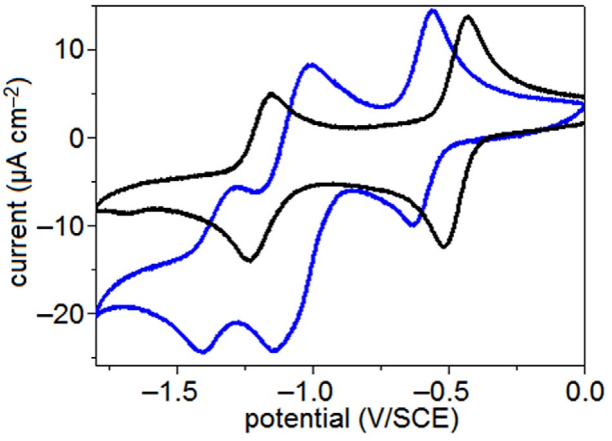



**Figure 2.** CV profiles of **1 g** (0.91 mM, blue) and **3 h** (1.19 mM, black), measured in DMSO with N($n$Bu)$_4$PF$_6$ (0.1 M) electrolyte support at 25 °C. $v$ = 200 mV s$^{-1}$; reference electrode: KCl(3 M)/Ag/AgCl; working electrode: glassy carbon disk with 0.07 cm$^2$ area. A third redox wave, attributed to the nitro group, can be observed between the two one-electron transfers of 1,4-NQ.

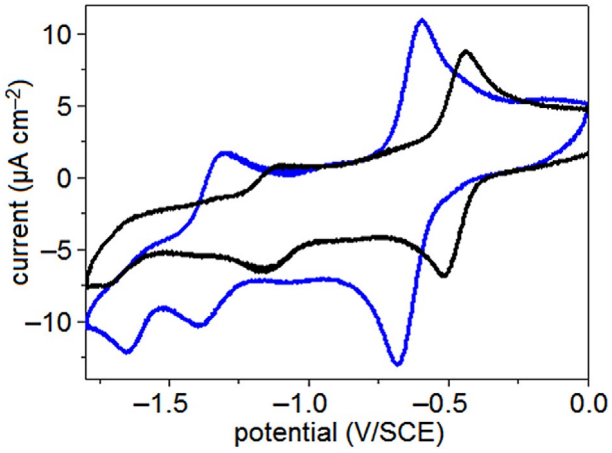



**Figure 3.** CV profiles of **1 i** (1.05 mM, black) and **3 i** (0.99 mM, blue), measured in DMSO with N($n$Bu)$_4$PF$_6$ (0.1 M) electrolyte support at 25 °C. $v$ = 200 mV s$^{-1}$; reference electrode: KCl(3 M)/Ag/AgCl; working electrode: glassy carbon disk with 0.07 cm$^2$ area. A third redox wave, attributed to the benzoyl carbonyl group, can be observed at more negative values.

(Table S3) of the second reduction step ($i_{pc2}$) leading to 1,4-NQ$^{2-}$ were systematically found to be weaker with respect to the reduction step ($i_{pc1}$) leading to the monoradical anion 1,4-NQ$^{\bullet-}$. This electrochemical behavior may result from comproportionation reactions ($E^1_{1/2}$-$E^2_{1/2}$ ~0.65–0.74 V and ln $K_{comp}$ ranging from ~23 to 26)[30–33] and/or by fast and irreversible dimerization[33] between 1,4-NQ$^{2-}$ and 1,4-NQ to afford the electro-inactive dimer (1,4-NQ)$_2^{2-}$.

With the exception of the nitro-containing compound, a clear relationship (Figure 4) was observed between the half-wave potentials of the first ($E^1_{1/2}$) and second ($E^2_{1/2}$) electrochemical processes. This feature strongly suggests that the substituents on the benz(o)yl core induce the same electronic effects on both the 1,4-NQ unit and its one-electron-reduced semiquinone, 1,4-NQ$^{\bullet-}$. However, exceptions were observed for benzoylmenadiones **3 o** (steric effect of the 2′-F substituent) and **3 q**/**3 s** (steric effect of the 2′-OCH$_3$ substituent), nitro-containing benz(o)ylmenadiones **1 g** and **3 g** (formation of a one-electron-reduced NO$_2$ intermediate), and benz(o)yl-plumbagin derivatives **2 a** and **4 a** (strong hydrogen bond between the 5-OH group and the 1,4-NQ carbonyl group).

To gain insight into the influence of the benz(o)yl substituents on the electrochemical properties of the 1,4-NQ core, the Hammett[34,35] approach (Figure 5) was used, as previously done for simpler menadiones.[22] The 1,4-NQ reactivity can be anodically or cathodically shifted by adding electron-releasing

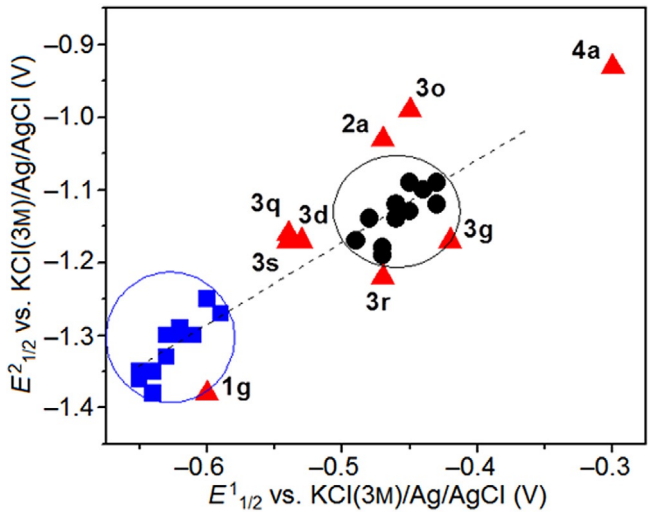



**Figure 4.** Variation in $E^2_{1/2}$ (V, second redox step) as a function of $E^1_{1/2}$ (V, first redox step). $v = 200$ mV s$^{-1}$; reference electrode: KCl(3 M)/Ag/AgCl; working electrode: glassy carbon disk with 0.07 cm$^2$ area; auxiliary electrode: Pt wire. The dashed line is only provided as a guide. ■ 3-Benzylmenadiones; ● 3-benzoylmenadiones; ▲ compound with particular behavior (see text).

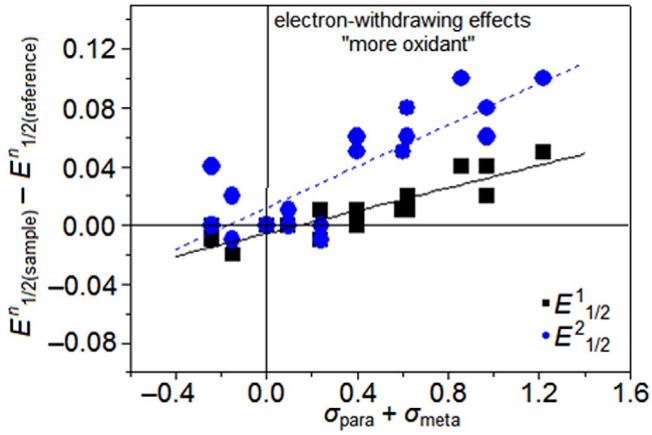



**Figure 5.** Plots of $\Delta E^1_{1/2}$ (■) and $\Delta E^2_{1/2}$ (●) vs. $\Sigma(\sigma_m + \sigma_p)$ for the substituted benz(o)ylmenadiones series. Solvent: DMSO; $I = 0.1$ M N($n$Bu)$_4$PF$_6$; $T = 25$ °C; $v = 200$ mV s$^{-1}$; reference electrode: KCl(3 M)/Ag/AgCl; working electrode: glassy carbon disk of 0.07 cm$^2$ area; auxiliary electrode: Pt wire. $\Delta E^1_{1/2} = 0.040(5) \times (\Sigma(\sigma_m + \sigma_p)) - 0.006(3)$ ($R^2 = 0.820$); $\Delta E^2_{1/2} = 0.09(1) \times (\Sigma(\sigma_m + \sigma_p)) - 0.003(6)$ ($R^2 = 0.815$). For the 4′-substituted systems, we used the sum $(\sigma_m + \sigma_p)$ to describe the electronic effect of a substituent. In the case of polysubstituted menadiones, the contribution of each substituent was considered.

or -withdrawing substituents either to the electrophore[22] or to the benz(o)yl moiety (Table 2). The half-wave potentials ($E^n_{1/2(reference)}$, $n = 1$ or 2) of the references (**1 h** for the benzyl series and **3 h** for the benzoyl series) were shifted to new values ($E^n_{1/2(sample)}$, $n = 1$ or 2) by introducing one or several substituents (Table 2 and Figure 5). With the exception of peculiar systems (nitro derivatives **1 g** and **3 g** and plumbagin analogues **2 a** and **4 a**) and those inducing steric interactions (2′-substituted-3-benz(o)ylmenadiones were not considered, Figure 5), the potential shifts could be theoretically predicted by the Hammett-Zuman relationship. Irrespective of the benzyl or benzoyl series, Figure 5 illustrates the variation of $\Delta E^n_{1/2}$ ($n = 1$ or 2) and demonstrates a linear dependence of $\Delta E^1_{1/2}$ and $\Delta E^2_{1/2}$ with the electronic substituent constants ($\Sigma(\sigma_m + \sigma_p)$). This feature indicates that substituents borne by the benz(o)yl moiety electronically influence the 1,4-NQ core in a related manner.

Overall, for 3-benz(o)ylmenadiones, the reaction constant ($\rho_R$) for the first electron transfer (Figure 5) was calculated to be 0.68(1). This value was significantly lower than those measured for the diversely substituted menadiones ($\rho^1_R = 2.20$),[22] demonstrating that substitution on the western region (i.e., 1,4-NQ core) of the benz(o)ylmenadiones has a much stronger effect on the oxidant character than does substitution on the eastern side (i.e., benz(o)yl unit). As electron delocalization through conjugation can be excluded for the 3-benz(o)ylmenadiones because the benzylic methylene or benzoyl carbonyl are not good electronic relays, other interactions, such as CH–π or dipole–dipole, can be proposed to explain the sensitivity of the 1,4-NQ electrophore to the benz(o)yl substitution pattern. In contrast with the 3-benz(o)ylmenadiones, their corresponding one-electron-reduced species are much more sensitive to substitution on the eastern side, as shown by the reaction constant ($\rho_R$) for the second electron transfer ($\rho^2_R = 1.5(2)$). This might be rationalized by other interactions (e.g., anion–π···) that take place within the negatively charged one-electron-reduced species. As a general rule, and regardless of the nature of the substitution and its position (east vs. west), electron-withdrawing substituents decrease the electronic density of the 1,4-NQ electroactive unit and facilitate the reduction of these compounds, while electron-donating substituents (Me, *t*Bu) lessen the propensity of the corresponding redox compounds to be reduced (Figure 5).

In the strict implementation of Hammett-Zuman equation, its intercept should be null. This is almost the case, within experimental errors, for $\Delta E^1_{1/2}$ and $\Delta E^2_{1/2}$ (Figure 5). This suggests that the two consecutive electron transfers are not complicated by coupled chemical reactions such as proton transfer.

Comparison of the 3-benzoylmenadiones (i.e., putative bioactivated metabolite) and the benzyl analogues (i.e., putatively acting as prodrug) shows that the benzylic oxidation believed to occur within the pRBCs plays a crucial role in electrochemical characteristics and may thereby explain the observed antimalarial activity. Both one-electron redox waves $E^1_{1/2}$ and $E^2_{1/2}$ (Figure 5) were substantially shifted to more positive potential values, meaning that the oxidized compounds at the benzylic position became more oxidized and therefore were more prone to transfer their electron to ferric targets such as metHb(Fe$^{III}$) (see below). With some exceptions that are discussed below, the redox potentials were anodically shifted by ~160 to 200 mV (see Table 4 below). This is clearly the result of the new substitution by a keto functionality that stabilizes the π-electrons of the menadione core (regardless of the oxidation state) by delocalization and conjugation. Deviation from this trend was observed for benzyl/benzoyl pairs displaying steric interactions (**1 q**/**3 q**) and for the phenolic/carboxylic-substituted systems (**1 d**/**3 d** and **1 f**/**3 f**), the nitro-substituted derivatives (**1 g**/**3 g**) and the plumbagin analogues (**2 a**/**4 a**). Steric interaction, intramolecular proton transfer within the reduced

species, and an additional redox pair centered on a nitro functionality or strong hydrogen bond are among the processes that can modulate the potential gap between the oxidized and reduced species (Table 2).

### QSPR modeling studies of the redox properties

Previous work[22] reported QSPR models for redox potential that were built on a set of various quinones and indolone-*N*-oxides, both novel and taken from literature. One of the explored modeling protocols, based on ISIDA molecular fragment counts (the best compromise between accuracy and technical web deployment costs) has been posted on our web server for public use. Basically, it predicts the redox potential values by adding fragment-specific increments for each of the key fragments shown, in the training stage, to best explain experimental property values. The approach accepts a structure file for organic compounds, then proceeds, for each molecule, to the detection and counting of the mentioned key fragments. Each occurrence of a key fragment triggers a fragment-specific increment, expressed in volts (which may be positive or negative, as calibrated by hand using training compounds) to be summed to the predicted redox potential value. Note that several other theoretical models based on different molecular descriptor schemes—notably Electronic Effect Descriptors (EED), designed for the purpose of modeling reactivity-related properties—have also been explored, with very promising results. Revisiting the technicalities of the previous modeling work is not the scope of the present paper. The reader is advised to refer to the previous article for details on the employed molecular descriptor schemes, which are also adopted in the present work.

Expansion of the chemical space of interest to include benzoyl derivatives naturally raises the question of the competence of the previous model with respect to this new chemotype, which has not been previously employed for training. Therefore, the first logical step was to challenge the old model to make a prediction for the newly synthesized compounds. The predictions by the model, are, on the absolute, quite inaccurate (root-mean-squared error [RMSE] = 0.176). This is not surprising, as the tested compounds are all derivatives of new families; the molecules contain benzyl and benzoyl substituents that were never present in the training set of the model, which had no chance to learn their impact on the redox potential value (Figure S43). The experimental, predicted (with and without the applicability domain), and corrected (by correction coefficients) values are available as Supporting Information (Table S4).

Table 3 reports the statistical parameters for new models built on a combination of old and new data with different methods. The training set for modeling consisted of 81 molecules (all previously used data and examples of both new families combined), and 14 benzyl and benzoyl derivatives were manually selected—or tested only after model building—to serve as an external test set.

The cross-validation $R^2$ scores were quite high for all modeling methods. Note that, in the evolutionary competition for the best descriptor space to host optimal SVM models, EED terms clearly outperformed individual descriptor spaces. However, models based on different ISIDA molecular fragments were combined into a consensus model (in which only models with $R^2$ of cross-validation $>0.5$ are accepted). This consensus effect, over many different ISIDA descriptor spaces, interestingly compensated for the advantage of EED over each individual ISIDA fragmentation scheme and also for the alleged advantage of nonlinear modeling. As shown, the single descriptor-space (EED) nonlinear model and the multi-fragmentation consensus approach eventually performed equally, and very well, in a threefold cross-validation challenge.

**Table 3.** Statistical characteristics of consensus multilinear regression (ISIDA MLR) and support vector machine (EED SVM) models for cross-validation (XV) and external test challenges.[a]

| Descriptor | Training (10×3−XV) | | Test | |
|---|---|---|---|---|
| | RMSE | $R^2$ | RMSE | $R^2$ |
| ISIDA MLR | 0.080 | 0.85 | 0.032 (0.027)[b] | 0.86 (0.90)[b] |
| EED SVM | 0.082 | 0.84 | 0.054 (0.030)[b] | 0.60 (0.88)[b] |

[a] RSME: root-mean-squared error; $R^2$: determination coefficient. [b] Adjusted statistical parameters when the outlier (**2a**) is excluded from the test set.

It is, nevertheless, of greater practical interest to evaluate a model's performance using the external test set. As Table 3 shows, consensus models based on ISIDA descriptors performed well in both cases, with the $R^2$ of the test being close to the $R^2$ of cross-validation. The model with EED descriptors had a globally lower $R^2$, but this was due to one molecule being predicted more poorly, which, due to the modest size of the test set, had a greater impact on the overall score (Figure 6). The poorly predicted molecule (**2a**) is the derivative with a hydroxy substituent, and the phenomenon has been reported previously.[22] However, the updated training set now includes examples of analogues with such an intramolecular hydrogen bond and, accordingly, the structural pattern was recognized when using ISIDA fragment counts. The monitored atom and bond sequences, with lengths ranging from two to ten, include the corresponding O=C–C:C–O pattern (":" represents an aromatic bond) of the fragment responsible for the intramolecular hydrogen bond and only appear in the context of the presence of a 5-hydroxy substituent. Therefore, the model learned to associate this fragment with a negative increment for the predicted redox potential (the hydrogen bond polarizes the quinone carbonyl even further, rendering the quinone system more electron-depleted). Of course, the presence of such a fragment itself is not sufficient to trigger a decrease in the redox potential; it must be determined whether the fragment is in the right context, that is, involving the actual quinone carbonyl and the adjacent phenolic hydroxy group in a hydrogen bonding position. Should the mentioned sequence appear in other moieties of a molecule, not connected to the quinone system, the model would also add the associated (negative) increment to the predicted redox potential and would likely result in an error. This shows that even the significantly extended compound set used in this work is still prone

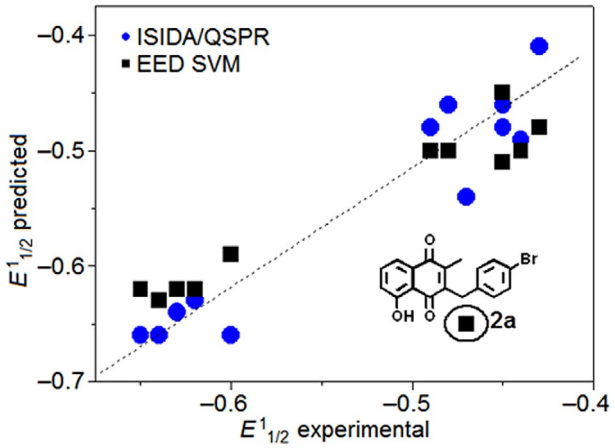



**Figure 6.** Plot of $E^1_{1/2}$ for test set compounds, experimental values vs. those predicted by all models: ISIDA MLR (●), predicted by ISIDA/QSPR software, and EED SVM (■), predicted by the web server application. The dotted line represents perfect fit. The outlier (benzyl derivative with hydroxy substituent in the naphthoquinone moiety) is highlighted.

to significant biases and sources of artifacts. The rule "presence of a O=C–C:C–O → presence of intramolecular hydrogen bond → lowered redox potential" was learned on the basis of training examples and also happened to hold for the test set compounds but was clearly not a general characteristic.

In contrast, EED descriptors do not explicitly consider connectivity patterns but only focus on through-bond electronic effects, ignoring all through-space non-bonded influences. As such, they cannot capture the atypical effect in 5-OH analogues, even if these were present in the training set. They fail to learn any specific rule for the 5-OH analogues and therefore commit an important error of +0.17 V when predicting the redox potential of **2a**. However, they are not prone to fail in the above-discussed scenario of external predictions for compounds with O=C–C:C–O moieties not responsible for redox behavior. If **2a** is discarded as an outlier, then RMSE = 0.030 V and $R^2 = 0.88$ (comparatively, for ISIDA descriptors RMSE = 0.027, $R^2 = 0.90$).

The comprehensive table of predictions for test set compounds by all methods is available as Supporting Information (Tables S4 and S5). SVM models based on EED descriptors are also available for users in a web application (http://infochim.u-strasbg.fr/webserv/VSEngine.html). When an SDF file is uploaded in the application, it automatically detects the redox center of the molecule. Evidently, the applicability domain of the model accepts only molecules with a carboxyl redox center, preferably a quinone or indolone moiety. Performance with other scaffolds of redox-active compounds is not guaranteed.

### Reduction of methemoglobin to hemoglobin

To evaluate the potency of our compounds to reduce metHb($Fe^{III}$) to oxyHb($Fe^{II}$) and to assess their redox characteristics on reduction kinetics, we employed a reduction assay coupled to the hGR/NADPH system in vitro (under quasi-physiological conditions), which regenerated the reduced species of our substrates as long as NADPH was present. This assay was recently established as a relevant in vitro model in our laboratory.[4,7] The UV/visible absorption spectrum of metHb($Fe^{III}$) between 300 and 700 nm was characterized by a Soret band of the $Fe^{III}$ heme, whose maximum absorbance is centered at 405 nm and was used as a valuable spectroscopic probe. Upon metHb($Fe^{III}$) reduction by the enzymatically generated reduced species of the redox-active compounds, the occurrence of the oxyHb($Fe^{II}$) species was associated with a bathochromic shift in the absorption maximum from 405 to ~410 nm (now corresponding to a low spin hexacoordinated ferrous ion), as well as the formation of two new less intense absorptions at ~536 and ~576 nm. As an example, Figure 7 depicts the spectral absorption variation recorded for **1c**, **3b**, and **6a** under these experimental conditions. These compounds represent some of the putative metabolites series that can be generated within the pRBCs, namely the benzyl-, benzhydrol-, and benzoylmenadiones.

Kinetic data (pseudo-first-order rate constant, $k_{red}$, $s^{-1}$) processed for a large series of compounds are shown in Table 4, together with the $E^1_{1/2}$ potential values previously determined using CV. This important dataset allowed us to highlight interesting properties related to the mechanism of action of the antimalarial benz(o)ylmenadiones. First, the benzylmenadiones (Figure 7A), irrespective of their substitutions, were not reactive with respect to metHb($Fe^{III}$). This feature can be clearly related to their redox potentials, which are ~160 to 200 mV cathodically shifted with respect to their benzoyl analogues (Table 4).

Bioactivation through benzylic oxidation leading to a benzhydrol (e.g., **6a**, Figure 7C, Table 4) or a benzoyl (Figure 7B, Table 4) analogue markedly favors metHb($Fe^{III}$) reduction. The higher oxidant character of the benzhydrol- or benzoylmenadiones therefore appears to be a key parameter. Notably, the redox properties of the redox cycler must be subtly fine-tuned, as their redox potentials should be in a narrow range of potential values, defined by the GR(flavine FAD/$FADH_2$)/NADPH[36] and the metHb($Fe^{III}$)/Hb($Fe^{II}$).[37] Deviation from this prevents either reduction of the electrophore by the GR/NADPH pair or reduction of the metHb($Fe^{III}$) by the reduced electrophore. This feature is exemplified in Figure 8 by benzylmenadiones **1c** or **1u** whose $E^1_{1/2}$ values (~0.6 V) are close to that of menadione (Table 2) and are not proper for efficient reduction by the GR/NADPH pair. On the other hand, the system with the greatest oxidant effects examined so far in this extensive work (i.e., plumbagin analogue **4a**, $E^1_{1/2} = -0.30$ V) was far from being the most efficient system to rapidly target metHb($Fe^{III}$) in the reduction assay coupled to the hGR/NADPH system in vitro.

Even though a dependence of the rate constant of the metHb($Fe^{III}$) reduction as a function of the redox potential $E^1_{1/2}$ can be deduced as anticipated within a limited potential range as defined above, several systems deviate from this linear variation and were found to be very informative. As previously noted, plumbagin derivative **4a** deviates from this trend because of its particular structure and subsequent redox behavior (i.e., a strong hydrogen bond between the 5-OH group and one of the carbonyl groups that renders the compound too

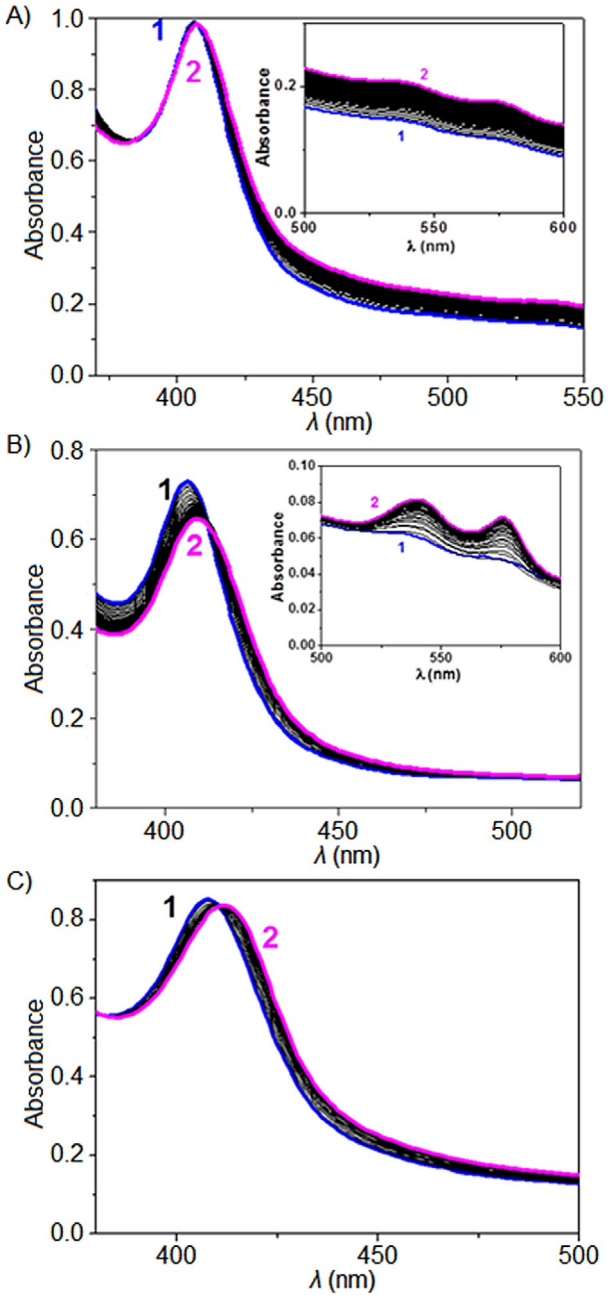



**Figure 7.** UV/visible absorption spectra recorded as a function of time in the coupled assay in the presence of the hGR/NADPH system. A) Spectral variations showing metHb($Fe^{III}$) reduction in the presence of benzylmenadione **1c**. B) Spectral variations showing metHb($Fe^{III}$) reduction in the presence of benzoylmenadione **3b**. C) Spectral variations showing metHb($Fe^{III}$) reduction in the presence of benzhydrolmenadione **6a**. Solvent: $H_2O$, phosphate buffer (47 mM) pH 6.9, EDTA (1 mM), KCl (200 mM); $T = 25.0\,^{\circ}C$; NADPH (120 μM) + hGR (88 nM) + metHb (8 μM) + GSSG (20 μM) + substrate (40 μM); (1) $t = 0$; (2) $t = 120$ min.

**Table 4.** Electrochemical data (first redox wave), measured by cyclic voltammetry (CV)[a] and kinetic data, related to the reduction of methemoglobin in the hGR/NADPH coupled assay.[b]

| Compd | $E^1_{1/2}$ [V] | $k_{red}$ [$s^{-1}$] | Compd | $E^1_{1/2}$ [V] | $k_{red}$ [$10^{-4}\,s^{-1}$] | $\Delta E^1_{1/2}$ [mV] |
|---|---|---|---|---|---|---|
| **1a** | −0.62 | – | **3a** | −0.46 | 5.0(3) | 160 |
| **1b** | −0.63 | – | **3b** | −0.46 | 7.7(3) | 170 |
| **1c** | −0.59 | nr | **3c**[4] | −0.43 | 13.1 | 160 |
| **1d** | −0.65 | – | **3d** | −0.53 | 101(2) | 120 |
| **1e** | −0.64 | – | **3e** | −0.47 | 14.2(4) | 170 |
| **1f** | −0.65 | – | **3f** | −0.49 | 36.2(8) | 160 |
| **1g** | −0.60 | – | **3g** | −0.42 | 39.0(6) | 180 |
| **1h** | −0.63 | – | **3h** | −0.47 | – | 160 |
| **1i** | −0.64 | – | **3i** | −0.48 | – | 160 |
| **1j** | −0.62 | – | **3j** | −0.45 | 17(2) | 170 |
| | | | **3k** | −0.43 | 54(2) | |
| | | | **3l** | −0.49 | 11.4(3) | |
| | | | **3m** | −0.45 | 82(3) | |
| | | | **3n** | −0.44 | 16.6(5) | |
| | | | **3o** | −0.45 | – | |
| **1q** | −0.65 | – | **3q** | −0.54 | 15.1(4) | 110 |
| **1r** | −0.65 | – | **3r** | −0.47 | 9.0(3) | 180 |
| | | | **3s** | −0.54 | 15.0(3) | |
| | | | **3t** | −0.45 | 26(3) | |
| **1u** | −0.60 | nr | | | | |
| **2a** | −0.47 | – | **4a** | −0.30 | 23(1) | 170 |
| | | | **6a** | – | 33(1) | |

[a] $v = 200\ mV\,s^{-1}$; reference electrode: KCl(3 M)/Ag/AgCl; working electrode: glassy carbon disk of 0.07 $cm^2$ area; auxiliary electrode: Pt wire. Compounds used for the validation test shown in blue. [b] Solvent: $H_2O$, phosphate buffer (47 mM) pH 6.9, EDTA (1 mM), KCl (200 mM); $T = 25.0\,^{\circ}C$; NADPH (120 μM) + hGR (80 nM) + metHb (8 μM) + GSSG (20 μM) + substrate (40 μM). nr: no reaction; –: not measured.

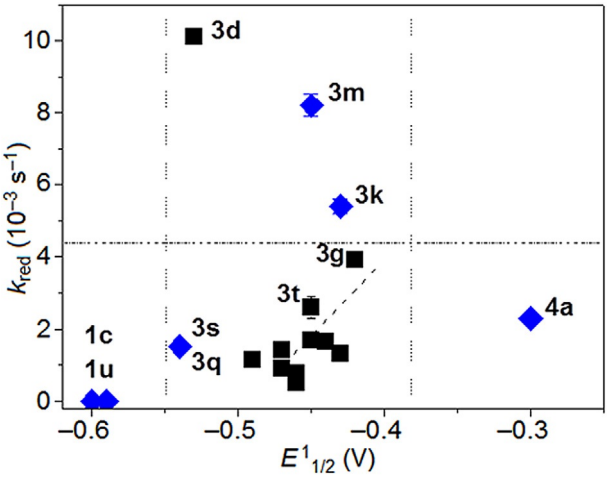



**Figure 8.** Schematic representation of the influence of the redox potential $E^1_{1/2}$ (V) on the kinetics of metHb($Fe^{III}$) reduction. The dashed line is provided only as a guide.

highly oxidizing, that is, the $E^1_{1/2}$ redox potential is anodically shifted to −0.30 V). Likewise, the benzylmenadiones, which are significantly poorer oxidizing agents than the benzoyl analogues, are not reactive, because their redox properties are outside the potential range of efficacy. Compounds **3s** and **3r** behave differently because of steric interactions of the benzylic 2′-methoxy substitution. This is confirmed by derivative **3t**, which bears a less bulky 2′-fluorine substitution and follows the observed trend (Figure 8). Surprisingly, the two derivatives bearing a cyano group on the benzoyl unit (**3m** and **3k**) were found to be very reactive in the reduction assay. This would suggest an unexpected role for the cyano group that signifi-

cantly boosts the rate of metHb reduction. It should be mentioned that the electron transfer between the reduced benzoylmenadione and the ferric target might occur either through an inner (the involved redox sites interact) or outer (the involved redox sites remain separate before, during, and after the electron transfer event) sphere mechanism. This would suggest that the cyano group might interact with the ferric reactive center (or, to a lesser extent, with the globin) and favor the reduction process. Compound **3 d**, bearing a 4′-hydroxy substituent on the benzoyl unit, was found to be the most reactive system ($k_{red} = 1.01 \times 10^{-2}\,s^{-1}$) among the series examined and supports the idea that, in addition to the intrinsic redox character of the compound, its ability to interact with one of the putative ferric targets is also of importance to the global antimalarial activity. Indeed, we previously showed that some of the putative metabolites of **1 c** (i.e., plasmodione) efficiently prevent formation of β-hematin (i.e., synthetic hemozoin), which is one of the main mechanism by which *P. falciparum* escapes heme toxicity after methemoglobin digestion.

## Conclusions

The electrochemical properties of a series of polysubstituted 3-benz(o)ylmenadiones acting as efficient redox cyclers toward pRBCs were thoroughly investigated. This homogeneous set of compounds supplemented a preliminary work, focused on the evaluation of the steric/electronic effects of diverse substitution on the western part of the 3-benz(o)ylmenadiones, that is, the 1,4-NQ moiety. All of these compounds were observed to display a common pattern of two successive one-electron reversible transfers, separated by a potential gap, ranging from 0.6 to 0.8 V. The potential values of both electron transfers were shown to be drastically altered by the nature and/or the position of the substituents. We clearly identified the substituents and their subsequent electronic/steric/mesomeric properties that control the oxidant character of the electroactive menadione derivatives. The most important effects were by far those brought by the 1,4-NQ substitution, while those centered on the benz(o)yl unit were less critical but, however, of significance in the context of their redox cycling capacities. Importantly, this comprehensive study allowed us to evaluate the impact of the substitution (position and nature) on the redox properties of a promising class of multitargeted antimalarial drugs and will undoubtedly pave the way toward better drug design.

We also demonstrated that bioactivation of the drugs through putative benzylic oxidation markedly influence the oxidant character of the one- or two-electron-reduced species through a significant anodic shift in their redox potentials (Table 4). This accounts for the new keto functionality that increased delocalization and conjugation of the π-electrons of the menadione core, regardless of its oxidation state. The antimalarial efficiency seemingly results from both the fine-tuned redox properties of the drugs and the bioactivation (i.e., benzylic oxidation among other possibilities) with respect to several endogenous crucial partners, such as disulfide reductase and ferric targets. Developed QSPR models are able to predict the redox potential of various quinones and indolone-*N*-oxides with reasonable accuracy, and can therefore be used for the computer-aided design of new antimalarial agents. We assessed the potency of the drug candidates by evaluating the electron transfer efficiencies and kinetics between the reduced 3-benz(o)ylmenadiones generated by reduction by hGR under excess NADPH cofactor and methemoglobin metHb($Fe^{III}$). The 3-benzylmenadiones were not reactive, while the benzoyl analogues efficiently reduced metHb($Fe^{III}$). Even though a relationship can be drawn between the $E^{1}_{1/2}$ potential values and the apparent metHb($Fe^{III}$) reduction rate constants, the most potent systems were those displaying a cyano or hydroxy group on the benzoyl position that was thought to interact with the ferric center through an inner-sphere electron-transfer mechanism.

## Experimental Section

### *Physicochemical characterization*

**Solvents and materials**: Electrochemical and absorption spectroscopic properties of substituted 3-benz(o)ylmenadiones (Figures S1 and S2) and plumbagin (5-hydroxymenadione) were examined in spectroscopic grade DMSO (>99.9%, Sigma–Aldrich). All stock solutions were prepared by weighing solid products using a Mettler Toledo XA105 Dual Range balance. The complete dissolution of the ligands was obtained using an ultrasonic bath (Bandelin Sonorex RK102).

**Electrochemistry**: The redox potentials of the substituted 3-benz(o)ylmenadiones and plumbagin were measured by cyclic voltammetry (CV) and square-wave voltammetry (SWV) in DMSO (Table 1). DMSO can be classified as a dipolar aprotic protophilic solvent and was assumed as an suitable solvent[38,39,40] to characterize the electrochemical properties of our menadione analogues (measurable potential limits of +0.9 V to −3.9 V with respect to $F_c/F_c^+$, $E = +0.524$ V/Ag/AgCl).[39] CV and SWV voltammograms were recorded at room temperature (23±1 °C) in DMSO with 100 mM tetra-*n*-butylammonium hexafluorophosphate [$N(nBu)_4PF_6$] as a supporting and inert electrolyte.[41] CVs of the menadione or plumbagin derivatives (~$10^{-3}$ M) were first performed using a Voltalab 50 potentiostat/galvanostat (Radiometer Analytical MDE15 polarographic stand, PST050 analytical voltammetry, and CTV101 speed control unit), controlled by the Voltamaster 4 electrochemical software. A conventional three-electrode cell (10 mL) was employed in our experiments with a glassy carbon disk (GC, $s = 0.07\,cm^2$) set in a Teflon rotating tube as a working electrode, a platinum (Pt) wire as a counter electrode, and KCl(3 M)/Ag/AgCl reference electrode (+210 mV vs. NHE).[42] Prior to each measurement, the surface of the GC electrode was carefully polished with a 0.3 μm aluminum oxide suspension (ESCIL) on a silicon carbide abrasive sheet of 800/2400 grit. The GC electrode was then copiously washed with water and dried with paper towels and argon. The electrode was installed into the voltammetry cell along with a Pt wire counter electrode and the reference. Solutions containing the menadione or plumbagin derivatives were vigorously stirred and purged with $O_2$-free argon (Sigma Oxiclear cartridge) for 15 min before the voltammetry experiments were initiated and maintained under an argon atmosphere during measurement. The voltage sweep rate was varied from 50 to 300 $mV\,s^{-1}$, and several cyclic voltammograms were recorded from +0.5 V to −2.2 V. Peak potentials were measured at a scan rate of 200 $mV\,s^{-1}$ unless otherwise indicated.

Redox potentials were determined from oxidation and reduction potentials. Secondly, SWV voltammograms were recorded using the same analytical apparatus. A series of optimization studies were performed using the SWV parameters:[43] optimized values for these parameters were: pulse height ($\Delta E_p$), 50 mV; pulse width ($t_{step}$), 50 ms; step potential ($E_{step}$), 5 mV, step amplitude, 1 mV; scan rate, 20 mV $s^{-1}$; start, +0.5 V, end, −2.2 V.

**Methemoglobin reduction hGR/NADPH coupled assay**: Stock solutions of 2 mM of the drugs in DMSO and of 0.4 mM methemoglobin (from human hemoglobin, Sigma–Aldrich) were freshly prepared in water. Stock solutions of ~4 mM NADPH and 1 mM GSSG (oxidized L-glutathione, Serva) were prepared daily in hGR buffer ($KH_2PO_4$ [2.79 g], $K_2HPO_4{\cdot}3H_2O$ [6.04 g], EDTA [0.372 g], and KCl [14.91 g] in 1 L $H_2O$, pH 6.9). The pH was adjusted by dropwise addition of KOH (5 M). The final NADPH concentration was calculated from the absorbance recorded in the optical cell at 340 nm ($\varepsilon_{340}$ = 6.22 $mm^{-1}cm^{-1}$). All the solutions were kept at 0 °C during the experiments.

Recombinant human glutathione reductase (hGR) enzyme was purified and quantified (stock I, 0.124 mM), as described previously.[44,45] A hGR stock II solution was then prepared by dilution (1/10) of the enzyme stock solution with hGR buffer (pH 6.9). In a Hellma optical cell ($l$ = 1 cm), 902.9 μL of hGR buffer, 20 μL of methemoglobin stock solution (final concentration of 8 μM), 20 μL of drug stock solution (40 μM final concentration) and ~30 μL of NADPH solution (120 μM final concentration) were pre-incubated in the cell compartment of the absorption spectrophotometer at 25 °C (Varian Cary Dual Cell Peltier thermostat). Then, 7.1 μL of the hGR stock II solution (88 nM final concentration) was added to initiate the reaction, which was monitored by UV/Vis absorption spectrophotometry (200–800 nm) on an Agilent Cary 5000 spectrophotometer using the Scanning Kinetics method. A spectrum was taken every 2 min between 300 and 650 nm over 2 h. For fast kinetics, the spectral window was limited to the Soret band of methemoglobin (380–500 nm) and a spectrum was taken every 18 s over 30 min. The kinetic data were processed with the Specfit program.[46–51]

### *Datasets and modeling*

This section provides an overview of the data preparation and modeling workflow used in this work. The dataset described in our previous work[22,52] was enriched by novel in-house compounds and now consists of various benzo-, naphtho-, anthraquinones,[21] and indolone-*N*-oxides[53] taken from literature, as well as previously synthesized in-house menadione derivatives[22] and novel compounds described in this work. The equivalence of experimental measurement conditions was established. In total, we selected 95 compounds measured under similar experimental conditions. The structures were standardized (aromatization and charge-separated representation of nitro groups) prior to modeling using the ChemAxon Standardizer tool, version 14.10.20 (http://www.chemaxon.org).

Two types of descriptors were used: Electronic Effect Descriptors (EED)[22] and ISIDA fragment descriptors.[54] EED characterize the global impact of the chemical environment on the reactivity of a given key atom (*K*) in a molecule. They are computed as simple additive topological distance-weighted contributions of each of the nitrogen atoms of the compound. Here, the carbon atom in the carboxyl group of the quinone or indolone moiety was chosen as a key atom. In the case of the quinone ring, two carboxyl groups are present, so the descriptors are calculated as an average value between the two.

Second, ISIDA fragment count descriptors[55] that encode a molecular structure by the vector constituted occurrences of the fragments of each type. They offer several advantages over other types of descriptors: versatility, simplicity and rapidity of calculations, various applicability domain schemes[55,56] (fragment control and bounding box strategies), and full support from our web server Predictor application (http://infochim.u-strasbg.fr/cgi-bin/predictor.cgi), which allows them to be easily accessible by general public. With that in mind, 44 different fragmentation schemes, representing atom and bond sequences in different lengths ranging from 2 to 10, were used for training and building a consensus model.

ISIDA/QSPR software[57] was used to build multilinear regression[58] models on the training set consisting of 81 molecules. The external test set contained 14 compounds that were either voluntarily kept aside or actually measured after model fitting. The software offers consensus model building with any number of fragmentation schemes, as well as variable selection and applicability domain monitoring during the model selection process.

In addition, given the significantly larger training set size with respect to the previous work, nonlinear modeling (SVM epsilon-regression) was also created, using the evolutionary *libsvm* tuning tool[59] for modeling. This ensures optimal selection of the descriptor space (of both ISIDA and EED descriptor candidates) and of SVM parameters yielding maximal robustness (i.e., best cross-validation performance) models. After optimization, EED descriptors were selected as the most powerful and were used henceforth in the SVM model.

The models were built by both methods following a three-fold cross-validation[60] procedure repeated ten times and were additionally evaluated on the designated external set. For ISIDA descriptors, a consensus model was applied with ISIDA/QSPR's FMF module. The top EED-based SVM model was deployed on the web server application (http://infochim.u-strasbg.fr/webserv/VSEngine.html) and then used for external validation.

## *Acknowledgements*

*The authors thank the International Center for Frontier Research in Chemistry (IC-FRC Innovation 2015 Program) in Strasbourg for supporting our project entitled, "Computer-Aided Design of Novel Antimalarial Naphthoquinones (CAD-NQ)", and for creating a proper framework for the scientific collaborations. This work was supported in part by the Centre National de la Recherche Scientifique (CNRS) and the University of Strasbourg (UMR 7509 CNRS-Unistra), Laboratoire d'Excellence (LabEx) ParaFrap (grant LabEx ParaFrap ANR-11-LABX-0024 to E.D.-C.). P.S. and A.V. thank the Program of Competitive Growth of the Kazan Federal University for support. E.D.-C. acknowledges former PhD students and co-workers Holger Bauer (**3h**), Laure Johann (**3o**), and Karène Urgin (**3r–3t**) for compound preparation. The New York University Abu Dhabi Undergraduate Research Program is gratefully acknowledged for providing summer research funding to I.D.*

**Keywords:** chemoinformatics · menadione · multitarget drugs · QSPR · redox potential

[1] D. Reker, M. Seet, M. Pillong, C. P. Koch, P. Schneider, M. C. Witschel, M. Rottmann, C. Freymond, R. Brun, B. Schweizer, B. Illarionov, A. Bacher, M. Fischer, F. Diederich, G. Schneider, *Angew. Chem. Int. Ed.* **2014**, *53*, 7079 – 7084; *Angew. Chem.* **2014**, *126*, 7199 – 7204.
[2] O. Blank, E. Davioud-Charvet, M. Elhabiri, *Antioxid. Redox Signaling* **2012**, *17*, 544 – 554.
[3] K. Ehrhardt, E. Davioud-Charvet, H. Ke, A. B. Vaidya, M. Lanzer, M. Deponte, *Antimicrob. Agents Chemother.* **2013**, *57*, 2114 – 2120.
[4] T. Müller, L. Johann, B. Jannack, M. Brückner, D. A. Lanfranchi, H. Bauer, C. Sanchez, V. Yardley, C. Deregnaucourt, J. Schrével, M. Lanzer, R. H. Schirmer, E. Davioud-Charvet, *J. Am. Chem. Soc.* **2011**, *133*, 11557 – 11571.
[5] M. Bielitza, D. Belorgey, K. Ehrhardt, L. Johann, D. A. Lanfranchi, V. Gallo, E. Schwarzer, F. Mohring, E. Jortzik, D. L. Williams, K. Becker, P. Arese, M. Elhabiri, E. Davioud-Charvet, *Antioxid. Redox Signaling* **2015**, *22*, 1337 – 1351.
[6] D. Belorgey, D. A. Lanfranchi, E. Davioud-Charvet, *Curr. Pharm. Des.* **2013**, *19*, 2512 – 2528.
[7] L. Johann, D. A. Lanfranchi, E. Davioud-Charvet, M. Elhabiri, *Curr. Pharm. Des.* **2012**, *18*, 3539 – 3566.
[8] K. Ehrhardt, C. Deregnaucourt, A.-A. Goetz, T. Tzanova, B. Pradines, S. H. Adjalley, S. Blandin, D. Bagrel, M. Lanzer, E. Davioud-Charvet, *unpublished results*.
[9] C. Hansch, A. Leo, *Exploring QSAR: Fundamentals and Applications in Chemistry and Biology*, American Chemical Society, Washington DC, **1995**.
[10] A. Tropsha, P. Gramatica, V. K. Gombar, *QSAR Comb. Sci.* **2003**, *22*, 69 – 77.
[11] R. Todeschini, V. Consonni, R. Mannhold, H. Kubinyi, H. Timmermann, *Handbook of Molecular Descriptors*, Wiley, Weinheim, **2008**.
[12] F. Ruggiu, G. Marcou, A. Varnek, D. Horvath, *Mol. Inf.* **2010**, *29*, 855 – 868.
[13] A. Varnek, D. Fourches, D. Horvath, O. Klimchuk, C. Gaudin, P. Vayer, V. Solov′ev, F. Hoonakker, I. V. Tetko, G. Marcou, *Curr. Comput.-Aided Drug Des.* **2008**, *4*, 191 – 198.
[14] E. Alpaydin, *Introduction to Machine Learning*, MIT Press, Cambridge, **2004**.
[15] K. Alizadeh, M. Shamsipur, *J. Mol. Struct.* **2008**, *862*, 39 – 43.
[16] J. Cape, M. Bowman, D. Kramer, *Phytochemistry* **2006**, *67*, 1781 – 1788.
[17] M. Namazian, P. Norouzi, R. Ranjbar, *J. Mol. Struct.* **2003**, *625*, 235 – 241.
[18] M. Namazian, P. Norouzi, *J. Electroanal. Chem.* **2004**, *573*, 49 – 53.
[19] M. R. Hadjmohammadi, K. Kamel, P. Biparva, *J. Solution Chem.* **2011**, *40*, 224 – 230.
[20] P. Thanikaivelan, V. Subramanian, J. R. Rao, V. U. Nair, *Chem. Phys. Lett.* **2000**, *323*, 59 – 70.
[21] A. Behesti, P. Norouzi, M. R. Ganjali, *Int. J. Electrochem. Sci.* **2012**, *7*, 4811 – 4821.
[22] M. Elhabiri, P. Sidorov, E. Cesar-Rodo, G. Marcou, D. A. Lanfranchi, E. Davioud-Charvet, D. Horvath, A. Varnek, *Chem. Eur. J.* **2015**, *21*, 3415 – 3424.
[23] J. M. Anderson, J. K. Kochi, *J. Am. Chem. Soc.* **1970**, *92*, 1651 – 1659.
[24] a) N. Jacobsen, K. Torssell, *Liebigs Ann. Chem.* **1972**, *763*, 135 – 147; b) N. Jacobsen, K. Torssell, *Acta. Chem. Scand.* **1973**, *27*, 3211 – 3216.
[25] C. Biot, H. Bauer, R. H. Schirmer, E. Davioud-Charvet, *J. Med. Chem.* **2004**, *47*, 5972 – 5983.
[26] K. Blumenstiel, R. Schöneck, V. Yardley, S. L. Croft, R. L. Krauth-Siegel, *Biochem. Pharmacol.* **1999**, *58*, 1791 – 1799.
[27] a) O. Hammerich in *Organic Electrochemistry Revised and Expanded, 5th ed.* (Eds: O. Hammerich, B. Speiser), CRC, Boca Raton, Chap. 30, pp. 1149 – 1201; b) W. H. Smith, A. J. Bard, *J. Am. Chem. Soc.* **1975**, *97*, 5203 – 5210.
[28] a) J. E. B. Randles, *Faraday Soc. Trans.* **1948**, *44*, 327 – 338; b) A. Sevcik, *Coll. Czechoslovak Chem. Commun.* **1948**, *13*, 349 – 377.
[29] J. Mauzeroll, A. J. Bard, *Proc. Natl. Acad. Sci. USA* **2004**, *101*, 7862 – 7867.
[30] E. N. da Silva, Jr., M. de Moura, A. V. Pinto, M. Pinto, M. de Souza, A. J. Araujo, C. Pessoa, L. V. Costa-Lotufo, R. C. Montenegro, M. O. de Moraes, V. F. Ferreira, M. O. F. Goulart, *J. Braz. Chem. Soc.* **2009**, *20*, 635 – 643.
[31] A. J. Araújo, A. A. de Souza, E. N. da Silva, Jr., J. D. B. Marinho, M. de Moura, D. D. Rocha, M. C. Vasconcellos, C. O. Costa, C. Pessoa, M. O. de Moraes, V. F. Ferreira, F. C. de Abreu, A. V. Pinto, R. C. Montenegro, L. V. Costa-Lotufo, M. O. F. Goulart, *Toxicol. in Vitro* **2012**, *26*, 585 – 594.
[32] P. S. Guin, S. Das, P. C. Mandal, *Int. J. Electrochem. Sci.* **2008**, *3*, 1016 – 1028.
[33] D. O. Wipf, K. R. Wehmeyer, R. M. Wightman, *J. Org. Chem.* **1986**, *51*, 4760 – 4764.
[34] L. P. Hammett, *J. Am. Chem. Soc.* **1937**, *59*, 96 – 103.
[35] P. Zuman, *Substituent Effects in Organic Polarography*, Plenum Press, New York, **1967**.
[36] a) D. M. Veine, L. D. Arscott, C. H. Williams, Jr., *Biochemistry* **1998**, *37*, 15575 – 15582; b) "Flavoenzyme Structure and Function—Approaches Using Flavin Analogues": D. Edmondson, S. Ghisla, in *Methods in Molecular Biology, Vol. 131* (Eds.: S. K. Chapman, G. A. Reid), Humana, Totowa, **1999**, pp. 157 – 179.
[37] F. W. Scheller, N. Bistolas, S. Liu, M. Jänchen, M. Katterle, U. Wollenberger, *Adv. Colloid Interface Sci.* **2005**, *116*, 111 – 120.
[38] I. M. Kolthoff, *Treatise on Analytical Chemistry, Part 1, Theory and Practice, Thermal Methods*, Wiley, New York, **1993**.
[39] A. Ashnagar, J. M. Bruce, P. L. Dutton, R. C. Prince, *Biochim. Biophys. Acta Gen. Subj.* **1984**, *801*, 351 – 359.
[40] N. G. Tsierkezos, *J. Solution Chem.* **2007**, *36*, 289 – 302.
[41] K. Izutsu, *Electrochemistry in Nonaqueous Solutions*, Wiley-VCH, Weinheim, **2002**.
[42] D. T. Sawyer, A. Sobkowiak, J. L. Roberts, *Electrochemistry for Chemists*, Wiley, New York, **1995**.
[43] V. Mirceski, S. Komorsky-Lovric, M. Lovric, *Square-Wave Voltammetry: Theory and Application*, Springer, Berlin, **2007**.
[44] A. Nordhoff, U. S. Bucheler, D. Werner, R. H. Schirmer, *Biochemistry* **1993**, *32*, 4060 – 4066.
[45] P. M. Färber, L. D. Arscott, C. H. Williams, K. Becker, R. H. Schirmer, *FEBS Lett.* **1998**, *422*, 311 – 314.
[46] H. Gampp, M. Maeder, C. J. Meyer, A. D. Zuberbühler, *Talanta* **1985**, *32*, 95 – 101.
[47] F. J. C. Rossotti, H. S. Rossoti, R. J. Whewell, *J. Inorg. Nucl. Chem.* **1971**, *33*, 2051 – 2065.
[48] H. Gampp, M. Maeder, C. J. Meyer, A. D. Zuberbühler, *Talanta* **1985**, *32*, 257 – 264.
[49] H. Gampp, M. Maeder, C. J. Meyer, A. D. Zuberbühler, *Talanta* **1986**, *33*, 943 – 951.
[50] D. W. Marquardt, *J. Soc. Ind. Appl. Math.* **1963**, *11*, 431 – 441.
[51] M. Maeder, A. D. Zuberbühler, *Anal. Chem.* **1990**, *62*, 2220 – 2224.
[52] a) E. Davioud-Charvet, T. Müller, H. Bauer, R. H. Schirmer (Centre National de la Recherche Scientifique), Int. PCT Pub. No. WO 2009118327 A1 20091001, **2009**; b) E. Davioud-Charvet, D. A. Lanfranchi, L. Johann, D. L. Williams, E. C. Rodo (Centre National de la Recherche Scientifique), Int. PCT Pub. No. WO 2012131010 A1 20121004, **2012**.
[53] K. Reybier, T. H. Y. Nguyen, H. Ibrahim, P. Perio, A. Montrose, P. L. Fabre, F. Nepveu, *Biochemistry* **2012**, *88*, 57 – 64.
[54] A. Varnek, D. Fourches, F. Hoonakker, V. P. Solov′ev, *J. Comput.-Aided Mol. Des.* **2005**, *19*, 693 – 703.
[55] D. Horvath, G. Marcou, A. Varnek, *J. Chem. Inf. Model.* **2009**, *49*, 1762 – 1776.
[56] I. V. Tetko, I. Sushko, A. K. Pandey, H. Zhu, A. Tropsha, E. Papa, T. Oberg, R. Todeschini, D. Fourches, A. Varnek, *J. Chem. Inf. Model.* **2008**, *48*, 1733 – 1746.
[57] A. Varnek, D. Fourches, V. Solov′ev, O. Klimchuk, A. Ouadi, I. Billard, *Solvent Extr. Ion Exch.* **2007**, *25*, 433 – 462.
[58] C. Agostinelli, *J. Appl. Stat.* **2002**, *29*, 825 – 840.
[59] D. Horvath, J. B. Brown, G. Marcou, A. Varnek, *Challenges* **2014**, *5*, 450 – 472.
[60] D. M. Hawkins, S. C. Basak, D. Millsk, *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 579 – 586.

Received: January 6, 2016
Published online on March 7, 2016

# Chapter 8

# Software development

Currently available tools for the visualization of GTM [184] produce static pictures demonstrating the distribution of property or molecules over the map (as it can be seen on Figures 3.3, 3.5, 3.4). While these facilitate the analysis of the results, render them visible and relatively interpretable, and discard the need for manual treating of textual form of GTM output, still they have their disadvantages. As a common example, it happens often that some molecules may occupy the same point in the latent space, it becomes therefore difficult to analyze the results visually. The GTM visualizer implemented in the GUI version of ISIDA/GTM tool [184] overcomes the problem: the user may choose to show all neighbors of a selected point, including those at distance 0. However, in order to render the results of GTM models accessible to public, it is of great practical interest to create a web-based tool for map visualization, that will allow users to access the map, navigate it and interact with it through a web browser.

That challenge has been taken during the thesis. As a language for the client-side visualization, JavaScript has been chosen. JavaScript is one of the most simple, versatile and effective languages used to extend functionality in websites. What is more important, JavaScript community has created numerous libraries for many aspects that may be needed for web-page functionality and appearance modification.

**D3.js** [224] is a JavaScript library for manipulating documents based on data. D3 allows one to bind arbitrary data to a Document Object Model (DOM), and then apply data-driven transformations to the document. For example, a table of data may be transformed into an interactive SVG bar-chart or scatter plot. These data handling capabilities make D3.js very suitable for the creation of interactive GTM visualizer.

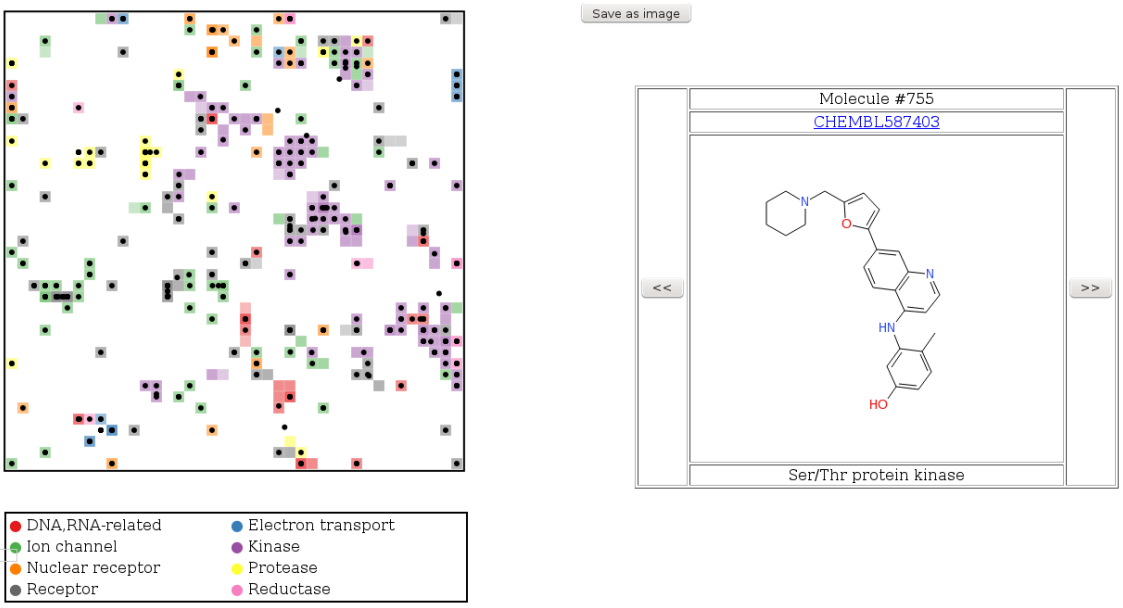



FIGURE 8.1: Prototype web-based GTM visualizer. The map on the left is built on TCAMS antimalarial dataset and colored by 8 classes of mechanisms. The right side contains the structure visualizer with the supplementary information on selected molecule.

## 8.1 GTM/Visualizer tool

Figure 8.1 shows the prototype web-based GTM/Visualizer tool created as a part of the thesis. First, let us describe the main functionality of the tool, then some technical details of its use will be given.

The tool is launched through an HTML page. The JavaScript code embedded in the page processes the given data and renders the map (on the left), its legend (below), and the molecule properties table (on the right).

The map demonstrates the distribution of molecules, drawn as black dots, and a property over the map. For the moment, only categorical properties (classes of compounds) with a maximum of 8 distinct categories are supported. The coloring of the map's nodes is based on the procedure described in section 3.4.1. For each node, the probability of every class is calculated, and the node is colored by the dominating class. Data density weighting scheme is also applied to indicate densely- and sparsely-populated zones on the map.

The legend is generated automatically from the given data.

The molecule property table helps the user to navigate the map. When one hovers over a molecule on the interactive map, all information that was given by the creator of the map on that molecule is shown there: chemical structure, ChEMBL ID, number in the

dataset, any supplementary fields such as class, target, *etc*. Arrow buttons are used to run through the dataset

### 8.1.1 Data input

For the script to build the interactive map, the following files are necessary:

- GTM output files. The most important file here is the responsibility matrix file. The responsibilities of molecules are used to calculate mean positions of molecules, probabilities of different classes and data density. In addition, latent space grid coordinates file helps the script to generate the nodes of the map. These files are generated by ISIDA/GTM tool, therefore, the user just has to load them.
- Custom file with the information on the dataset. The compounds should correspond to those whose responsibilities were given in the first file. It should be done in tab-separated table format, with the header which contains the names of properties. One of the columns should be named 'smiles' and should contain the structures of molecules in SMILES format, otherwise, the molecule visualizer will not work. The column 'class' contains the classes for each compound. It can be written as a name, the script will adapt itself to produce the legend and do the coloring accordingly to the number of unique class names. Obviously, error in class names will lead to erroneous coloring.

### 8.1.2 Coloring

As mentioned before, the color of nodes is determined by the responsibility distribution of molecules in these nodes. So, for each node, the responsibilities of compounds of each class corresponding to this point are summed up. The class that has the highest cumulative responsibility (so, the dominating class) in the one that gives the label to the node.

One of the most important features is the color intensity modulation according to the data density. This allows to delineate the applicability domain of the map: if a node is not populated (in other words, too few molecules contribute their responsibility to it), it cannot be robustly used for prediction, and it is thus not practical to color it. This density weighting is again based on the cumulative responsibility. The responsibilities of the whole dataset are summed up for each node and normalized to the range from 0 to 1. A threshold is given for the density, so that the nodes having inferior density value are shown as blank zones. In the current implementation, this threshold is set to $10^{-7}$.

### 8.1.3 Interactive features

The map on the left side of the page is interactive. The most basic interactive feature implemented in the tool is the demonstration of the structure of selected molecule and associated information in the table on the right. The chemical structure is visualized from the SMILES string by **openchemlib-js** library [225]. As we mentioned before, a problem that may arise during the visualization stage is that several molecules may be projected onto the same point in the 2D latent space. To overcome the problem, and further extend the functionality of chemical space analysis, a lasso plugin [226] has been added to the tool. It allows the user to define a custom area on the map, and for all molecules located in this area the IDs are output.

## 8.2 Future developments

This tool is nevertheless in the prototype stage of development. Among numerous features that we would like to implement, the most helpful for the end user would be the following:

1. **Projection of external set on the colored map**. For the moment, only one dataset is used for coloring of the map and positioning the molecules. While it is possible to change this behaviour by slightly modifying the code, of course, it's of utmost importance to implement this feature to allow the procedure for the end user.
2. **Acitivity landscape**. The current implementation supports only classification mode for the map. It is necessary to develop the regression counterpart of the landscape.
3. **Data management**. As the tool has been only tested on quite small datasets (up to 1000 compounds), its capabilities to manage large amounts of data have not been benchmarked. However, some conclusions can already be made based on the implementation details. All information on molecules is kept in form of a JavaScript object in RAM, which may be memory-consuming. A possible solution to the problem is to keep all the information as a database, and only access it when needed.
4. **Map customization**. Currently, most of map's parameters (such as the density threshold) are defined in the code. It should be possible for the end user to interactively change them.

# Conclusion

This thesis is dedicated to the concept of the analysis of chemical space, and the application of that concept to antimalarial compounds. Malaria is one of the most deadly tropical diseases, and the development of drug-resistant strains of the malarial parasite results in the rapid obsoletion of existing anti-Plasmodial drugs. The complexity of the problem is also linked to the lack of knowledge on the precise biological targets and modes of action of these drugs.

The analysis of the chemical space of antimalarial compounds here is done with the aid of the Generative Topographic Mapping (GTM) method. The GTM has been previously used for visualization and QSAR modeling of single-property datasets, however, in order to overcome the heterogeneity of the existing data, a concept of Universal maps is developed in discussed in detail in this thesis. These maps allow the simultaneous projection and prediction of several related or unrelated properties. As a proof-of-concept study, the universal maps of drug-like chemical space have been constructed. They encompass the ligands for more than 140 human proteins, and their "universality" have been demonstrated by building models for almost 500 bioactivity measures, including the antimalarial activity.

For the construction of a general map of the chemical space of antimalarial compounds, a dedicated dataset has been collected and curated. It contains 15000 compounds, out of which 1140 are annotated by a biological target or metabolic pathway linked to its antimalarial activity. Based on this set, three types of maps are built: local, global, and universal. All these maps preform well in predicting compounds active against the parasite. The main focus is, however, on their capabilities in the analysis of chemical space. The application of privileged responsibility patterns approach facilitates such analysis. With this approach, we have been able to study the overlap of data coming from different sources, detect *terra incognita* of the antimalarial chemical space, delineate zones corresponding to various mechanisms of action, as well as highlight the inconsistencies in target annotations for certain classes of compounds. Moreover, a tool for interactive map visualization and analysis has been developed.

In addition, QSPR models for the redox potential of naphthoquinones – a class of antimalarial compounds – have been developed. Interference with the redox cycle in the parasite is one of intensively studied mechanism of antimalarial activity. For this, a new type of descriptors has been proposed – Electronic Effects Descriptors (EED). The developed models have demonstrated very good performance.

We hope that the developed concepts and models will help medicinal chemists in the battle against malaria or other illnesses. Of course, many aspects of the *in silico* drug design could not fit in the frame of a thesis. An application of proposed tools for the virtual screening for new active antimalarial compounds is one of such facets. The prediction of activity or mode of action would require an extensive experimental support. Moreover, the work here is based on 2D descriptors. Introducing 3D structure in form of 3D QSAR, docking, or pharmacophore modeling, would be of great interest for the project, in our opinion. There is also always room for improvement, and the tools and concepts discussed in the thesis are bound to be ameliorated in the future.

# Bibliography


[1] World Health Organization, "World malaria report," *Geneva, Switzerland: World Health Organization*, 2014.

[2] K. I. Barnes, *Antimalarial Drugs and the Control and Elimination of Malaria*, pp. 1–17. Basel: Springer Basel, 2012.

[3] World Health Organization, "World malaria report," *Geneva, Switzerland: World Health Organization*, 2015.

[4] R. W. Snow, C. A. Guerra, A. M. Noor, H. Y. Myint, and S. I. Hay, "The global distribution of clinical episodes of plasmodium falciparum malaria," *Nature*, vol. 434, no. 7030, pp. 214–217, 2005.

[5] N. J. White, "The role of anti-malarial drugs in eliminating malaria," *Malaria journal*, vol. 7, no. 1, p. S8, 2008.

[6] A. Hott, M. S. Tucker, D. Casandra, K. Sparks, and D. E. Kyle, "Fitness of artemisinin-resistant Plasmodium falciparum in vitro," *Journal of Antimicrobial Chemotherapy*, vol. 70, no. 10, pp. 2787–2796, 2015.

[7] H. Liu, Y. Ding, L. A. Walker, and R. J. Doerksen, "Computational study on the effect of exocyclic substituents on the ionization potential of primaquine: Insights into the design of primaquine-based antimalarial drugs with less methemoglobin generation," *Chemical research in toxicology*, vol. 28, no. 2, pp. 169–174, 2015.

[8] O. Blank, E. Davioud-Charvet, and M. Elhabiri, "Interactions of the antimalarial drug methylene blue with methemoglobin and heme targets in Plasmodium falciparum: a physico-biochemical study," *Antioxidants & redox signaling*, vol. 17, no. 4, pp. 544–554, 2012.

[9] M. Jensen and H. Mehlhorn, "Seventy-five years of resochin® in the fight against malaria," *Parasitology research*, vol. 105, no. 3, p. 609, 2009.

[10] N. J. White, W. Pongtavornpinyo, R. J. Maude, S. Saralamba, R. Aguas, K. Stepniewska, S. J. Lee, A. M. Dondorp, L. J. White, and N. P. Day,

"Hyperparasitaemia and low dosing are an important source of anti-malarial drug resistance," *Malaria Journal*, vol. 8, no. 1, p. 253, 2009.

[11] World Health Organization, "World malaria report," *Geneva, Switzerland: World Health Organization*, 2010.

[12] A. V. S. Hill, "Vaccines against malaria," *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, vol. 366, no. 1579, pp. 2806–2814, 2011.

[13] D. G. Heppner, K. E. Kester, C. F. Ockenhouse, N. Tornieporth, O. Ofori, J. A. Lyon, V. A. Stewart, P. Dubois, D. E. Lanar, U. Krzych, P. Moris, E. Angov, J. F. Cummings, A. Leach, B. T. Hall, S. Dutta, R. Schwenk, C. Hillier, A. Barbosa, L. A. Ware, L. Nair, C. A. Darko, M. R. Withers, B. Ogutu, M. E. Polhemus, M. Fukuda, S. Pichyangkul, M. Gettyacamin, C. Diggs, L. Soisson, J. Milman, M.-C. Dubois, N. Garçon, K. Tucker, J. Wittes, C. V. Plowe, M. A. Thera, O. K. Duombo, M. G. Pau, J. Goudsmit, W. R. Ballou, and J. Cohen, "Towards an RTS,S-based, multi-stage, multi-antigen vaccine against falciparum malaria: progress at the Walter Reed Army Institute of Research," *Vaccine*, vol. 23, no. 17, pp. 2243 – 2250, 2005. Vaccines and Immunisation. Based on the Fourth World Congress on Vaccines and Immunisation.

[14] World Health Organization, "Malaria vaccine: WHO position paper," *Weekly epidemiological record*, no. 19, pp. 33–52, 2016.

[15] T. C. Luke and S. L. Hoffman, "Rationale and plans for developing a non-replicating, metabolically active, radiation-attenuated Plasmodium falciparum sporozoite vaccine," *Journal of Experimental Biology*, vol. 206, no. 21, pp. 3803–3808, 2003.

[16] Sanaria, "Sanaria®PfSPZ vaccine against malaria receives FDA fast track designation," *Sanaria Press-release*, 2016.

[17] E. A. Winzeler, "Malaria research in the post-genomic era," *Nature*, vol. 455, no. 7214, p. 751, 2008.

[18] E. L. Flannery, D. A. Fidock, and E. A. Winzeler, "Using genetic methods to define the targets of compounds with antimalarial activity: Miniperspectives series on phenotypic screening for antiinfective targets," *Journal of medicinal chemistry*, vol. 56, no. 20, pp. 7761–7771, 2013.

[19] D. Plouffe, A. Brinker, C. McNamara, K. Henson, N. Kato, K. Kuhen, A. Nagle, and F. Adrian, "In silico activity profiling reveals the mechanism of action of antimalarials discovered in a high-throughput screen," *Proc Natl Acad Sci*, vol. 410, pp. 9059–9064.

[20] "Life cycle of the malaria parasite." from Epidemiology of Infectious Diseases. Available at: http://ocw.jhsph.edu. Johns Hopkins Bloomberg School of Public Health. Creative Commons BY-NC-SA.

[21] P. M. O'Neill, V. E. Barton, S. A. Ward, and J. Chadwick, *4-Aminoquinolines: Chloroquine, Amodiaquine and Next-Generation Analogues*, pp. 19–44. Basel: Springer Basel, 2012.

[22] J. Petithory, "On the discovery of the parasite of malaria by A. Laveran: Bône 1878-Constantine 1880," 1994.

[23] G. Stork, D. Niu, R. A. Fujimoto, E. R. Koft, J. M. Balkovec, J. R. Tata, and G. R. Dake, "The first stereoselective total synthesis of quinine," *Journal of the American Chemical Society*, vol. 123, no. 35, pp. 8644–8644, 2001.

[24] F. Loeb, W. Clark, G. Coatney, L. Coggeshall, F. Dieuaide, A. Dochez, E. Hakansson, E. Marshall, C. Marvel, O. McCoy, *et al.*, "Activity of a new antimalarial agent, chloroquine (SN 7618): Statement Approved by the Board for Coordination of Malarial Studies," *Journal of the American Medical Association*, vol. 130, no. 16, pp. 1069–1070, 1946.

[25] P. A. Winstanley, S. A. Ward, and R. W. Snow, "Clinical status and implications of antimalarial drug resistance," *Microbes and infection*, vol. 4, no. 2, pp. 157–164, 2002.

[26] P. J. Rosenthal, *Antimalarial chemotherapy: mechanisms of action, resistance, and new directions in drug discovery*. Springer Science & Business Media, 2001.

[27] W. Peters *et al.*, "Chemotherapy and drug resistance in malaria.," *Chemotherapy and drug resistance in malaria.*, 1970.

[28] N. Surolia and G. Padmanaban, "Chloroquine inhibits heme-dependent protein synthesis in plasmodium falciparum," *Proceedings of the National Academy of Sciences*, vol. 88, no. 11, pp. 4786–4790, 1991.

[29] H. Ginsburg and T. G. Geary, "Current concepts and new ideas on the mechanism of action of quinoline-containing antimalarials," *Biochemical pharmacology*, vol. 36, no. 10, pp. 1567–1576, 1987.

[30] D. L. V. Jagt, L. A. Hunsaker, and N. M. Campos, "Characterization of a hemoglobin-degrading, low molecular weight protease from Plasmodium falciparum," *Molecular and biochemical parasitology*, vol. 18, no. 3, pp. 389–400, 1986.

[31] D. J. Sullivan, I. Y. Gluzman, D. G. Russell, and D. E. Goldberg, "On the molecular mechanism of chloroquine's antimalarial action," *Proceedings of the National Academy of Sciences*, vol. 93, no. 21, pp. 11865–11870, 1996.

[32] D. J. Sullivan, H. Matile, R. G. Ridley, and D. E. Goldberg, "A common mechanism for blockade of heme polymerization by antimalarial quinolines," *Journal of Biological Chemistry*, vol. 273, no. 47, pp. 31103–31107, 1998.

[33] C. K. Carney, L. Pasierb, and D. Wright, *Heme Detoxification in Malaria: A Target Rich Environment*, ch. 15, pp. 263–280.

[34] E.-M. Patzewitz, J. E. Salcedo-Sora, E. H. Wong, S. Sethia, P. A. Stocks, S. C. Maughan, J. A. Murray, S. Krishna, P. G. Bray, S. A. Ward, *et al.*, "Glutathione transport: a new role for pfcrt in chloroquine resistance," *Antioxidants & redox signaling*, vol. 19, no. 7, pp. 683–695, 2013.

[35] J. Wunderlich, P. Rohrbach, and J. P. Dalton, "The malaria digestive vacuole," *Frontiers in Bioscience*, vol. 4, pp. 1424–1448, 2012.

[36] S. Pagola, P. W. Stephens, D. S. Bohle, A. D. Kosar, and S. K. Madsen, "The structure of malaria pigment [beta]-haematin," *Nature*, vol. 404, pp. 307–310, Mar 2000.

[37] T. J. Egan, "Physico-chemical aspects of hemozoin (malaria pigment) structure and formation," *Journal of inorganic biochemistry*, vol. 91, no. 1, pp. 19–26, 2002.

[38] D. C. Warhurst, J. C. Craig, I. S. Adagu, D. J. Meyer, and S. Y. Lee, "The relationship of physico-chemical properties and structure to the differential antiplasmodial activity of the cinchona alkaloids," *Malaria Journal*, vol. 2, no. 1, p. 26, 2003.

[39] A. Slater, W. J. Swiggard, B. R. Orton, W. D. Flitter, D. E. Goldberg, A. Cerami, and G. B. Henderson, "An iron-carboxylate bond links the heme units of malaria pigment," *Proceedings of the National Academy of Sciences*, vol. 88, no. 2, pp. 325–329, 1991.

[40] A. B. S. Sidhu, D. Verdier-Pinard, and D. A. Fidock, "Chloroquine resistance in Plasmodium falciparum malaria parasites conferred by pfcrt mutations," *Science*, vol. 298, no. 5591, pp. 210–213, 2002.

[41] D. A. Fidock, T. Nomura, A. K. Talley, R. A. Cooper, S. M. Dzekunov, M. T. Ferdig, L. M. Ursos, B. Naudé, K. W. Deitsch, X.-z. Su, *et al.*, "Mutations in the P. falciparum digestive vacuole transmembrane protein PfCRT and evidence for

their role in chloroquine resistance," *Molecular cell*, vol. 6, no. 4, pp. 861–871, 2000.

[42] C. P. Sanchez, A. Rotmann, W. D. Stein, and M. Lanzer, "Polymorphisms within PfMDR1 alter the substrate specificity for anti-malarial drugs in Plasmodium falciparum," *Molecular microbiology*, vol. 70, no. 4, pp. 786–798, 2008.

[43] P. M. O'Neill, P. G. Bray, S. R. Hawley, S. A. Ward, and B. K. Park, "4-aminoquinolines – past, present, and future; a chemical perspective," *Pharmacology & therapeutics*, vol. 77, no. 1, pp. 29–58, 1998.

[44] E. B. Daily and C. L. Aquilante, "Cytochrome P450 2C8 pharmacogenetics: a review of clinical studies," *Pharmacogenomics*, vol. 10, no. 9, pp. 1489–1510, 2009.

[45] J. Ruscoe, M. Tingle, P. O'Neill, S. Ward, and B. Park, "Effect of disposition of mannich antimalarial agents on their pharmacology and toxicology," *Antimicrobial agents and chemotherapy*, vol. 42, no. 9, pp. 2410–2416, 1998.

[46] S. Auparakkitanon, S. Chapoomram, K. Kuaha, T. Chirachariyavej, and P. Wilairat, "Targeting of hematin by the antimalarial pyronaridine," *Antimicrobial agents and chemotherapy*, vol. 50, no. 6, pp. 2197–2200, 2006.

[47] R. N. Price, A.-C. Uhlemann, A. Brockman, R. McGready, E. Ashley, L. Phaipun, R. Patel, K. Laing, S. Looareesuwan, N. J. White, *et al.*, "Mefloquine resistance in Plasmodium falciparum and increased pfmdr1 gene copy number," *The Lancet*, vol. 364, no. 9432, pp. 438–447, 2004.

[48] T. J. Anderson, S. Nair, H. Qin, S. Singlam, A. Brockman, L. Paiphun, and F. Nosten, "Are transporter genes other than the chloroquine resistance locus (pfcrt) and multidrug resistance gene (pfmdr) associated with antimalarial drug resistance?," *Antimicrobial agents and chemotherapy*, vol. 49, no. 6, pp. 2180–2188, 2005.

[49] A. Nzila, "The past, present and future of antifolates in the treatment of Plasmodium falciparum infection," *Journal of Antimicrobial Chemotherapy*, vol. 57, no. 6, pp. 1043–1054, 2006.

[50] A. Nzila, S. A. Ward, K. Marsh, P. F. Sims, and J. E. Hyde, "Comparative folate metabolism in humans and malaria parasites (part I): pointers for malaria treatment from cancer chemotherapy," *Trends in parasitology*, vol. 21, no. 6, pp. 292–298, 2005.

[51] J. E. Hyde, "Targeting purine and pyrimidine metabolism in human apicomplexan parasites," *Current drug targets*, vol. 8, no. 1, pp. 31–47, 2007.

[52] F. Curd, D. Davey, and F. Rose, "Studies on synthetic antimalarial drugs: X. – some biguanide derivatives as new types of antimalarial substances with both therapeutic and causal prophylactic activity," *Annals of Tropical Medicine & Parasitology*, vol. 39, no. 3-4, pp. 208–216, 1945.

[53] H. Carrington, A. Crowther, D. Davey, A. Levi, and F. Rose, "A metabolite of 'Paludrine' with high antimalarial activity," *Nature*, vol. 168, no. 4286, pp. 1080–1080, 1951.

[54] T. Mutabingwa, C. Maxwell, I. G. Sia, F. Msuya, S. Mkongewa, S. Vannithone, J. Curtis, and C. Curtis, "A trial of proguanil-dapsone in comparison with sulfadoxine-pyrimethamine for the clearance of Plasmodium falciparum infections in tanzania," *Transactions of the Royal Society of Tropical Medicine and Hygiene*, vol. 95, no. 4, pp. 433–438, 2001.

[55] E. Falco, L. Goodwin, G. Hitchings, I. Rollo, and P. Russell, "2:4-diaminopyrimidines – a new series of antimalarials," *British Journal of Pharmacology*, vol. 6, no. 2, pp. 185–200, 1951.

[56] A. Nzila-Mounda, E. Mberu, C. Sibley, C. Plowe, P. Winstanley, and W. Watkins, "Kenyan Plasmodium falciparum field isolates: correlation between pyrimethamine and chlorcycloguanil activity in vitro and point mutations in the dihydrofolate reductase domain," *Antimicrobial agents and chemotherapy*, vol. 42, no. 1, pp. 164–169, 1998.

[57] R. Michel, "Comparative study of the association of sulfalene and pyrimethamine and of sulfalene alone in mass chemoprophylaxis of malaria," *Médecine tropicale: revue du Corps de santé colonial*, vol. 28, no. 4, p. 488, 1968.

[58] S. Bushby, "Combined antibacterial action in vitro of trimethoprim and sulphonamides. the in vitro nature of synergy.," *Postgraduate medical journal*, vol. 45, pp. Suppl–10, 1969.

[59] A. Gregson and C. V. Plowe, "Mechanisms of resistance of malaria parasites to antifolates," *Pharmacological reviews*, vol. 57, no. 1, pp. 117–145, 2005.

[60] A. Nzila, E. Mberu, P. Bray, G. Kokwaro, P. Winstanley, K. Marsh, and S. Ward, "Chemosensitization of Plasmodium falciparum by probenecid in vitro," *Antimicrobial agents and chemotherapy*, vol. 47, no. 7, pp. 2108–2112, 2003.

[61] O. Dar, M. Khan, and I. Adagu, "The potential use of methotrexate in the treatment of falciparum malaria: in vitro assays against sensitive and multidrug-resistant falciparum strains," *Japanese journal of infectious diseases*, vol. 61, no. 3, pp. 210–1, 2008.

[62] A. Nzila, J. Okombo, R. P. Becker, R. Chilengi, T. Lang, and T. Niehues, "Anticancer agents against malaria: time to revisit?," *Trends in parasitology*, vol. 26, no. 3, pp. 125–129, 2010.

[63] M. W. Mather, E. Darrouzet, M. Valkova-Valchanova, J. W. Cooley, M. T. McIntosh, F. Daldal, and A. B. Vaidya, "Uncovering the molecular mode of action of the antimalarial drug atovaquone using a bacterial system," *Journal of Biological Chemistry*, vol. 280, no. 29, pp. 27458–27465, 2005.

[64] H. J. Painter, J. M. Morrisey, and A. B. Vaidya, "Mitochondrial electron transport inhibition and viability of intraerythrocytic Plasmodium falciparum," *Antimicrobial agents and chemotherapy*, vol. 54, no. 12, pp. 5281–5287, 2010.

[65] M. A. Phillips, R. Gujjar, N. A. Malmquist, J. White, F. El Mazouni, J. Baldwin, and P. K. Rathod, "Triazolopyrimidine-based dihydroorotate dehydrogenase inhibitors with potent and selective activity against the malaria parasite Plasmodium falciparum," *Journal of medicinal chemistry*, vol. 51, no. 12, pp. 3649–3653, 2008.

[66] S. M. Ganesan, J. M. Morrisey, H. Ke, H. J. Painter, K. Laroiya, M. A. Phillips, P. K. Rathod, M. W. Mather, and A. B. Vaidya, "Yeast dihydroorotate dehydrogenase as a new selectable marker for Plasmodium falciparum transfection," *Molecular and biochemical parasitology*, vol. 177, no. 1, pp. 29–34, 2011.

[67] I. K. Srivastava, J. M. Morrisey, E. Darrouzet, F. Daldal, and A. B. Vaidya, "Resistance mutations reveal the atovaquone-binding domain of cytochrome b in malaria parasites," *Molecular microbiology*, vol. 33, no. 4, pp. 704–711, 1999.

[68] S. Looareesuwan, J. D. Chulay, C. J. Canfield, and D. Hutchinson, "Malarone (atovaquone and proguanil hydrochloride): a review of its clinical development for treatment of malaria. malarone clinical trials study group.," *The American journal of tropical medicine and hygiene*, vol. 60, no. 4, pp. 533–541, 1999.

[69] H. A. Karunajeewa, "Artemisinins: artemisinin, dihydroartemisinin, artemether and artesunate," in *Treatment and Prevention of Malaria*, pp. 157–190, Springer, 2011.

[70] World Health Organization, "Guidelines for the treatment of malaria," *WHO, Geneva*, 2010.

[71] X.-z. Su and L. H. Miller, "The discovery of artemisinin and Nobel Prize in Physiology or Medicine," *Science China. Life sciences*, vol. 58, no. 11, p. 1175, 2015.

[72] J. Li and B. Zhou, "Biological actions of artemisinin: insights from medicinal chemistry studies," *Molecules*, vol. 15, no. 3, pp. 1378–1397, 2010.

[73] S. Krishna, L. Bustamante, R. K. Haynes, and H. M. Staines, "Artemisinins: their growing importance in medicine," *Trends in pharmacological sciences*, vol. 29, no. 10, pp. 520–527, 2008.

[74] P. M. O'Neill, V. E. Barton, and S. A. Ward, "The molecular mechanism of action of artemisinin – the debate continues," *Molecules*, vol. 15, no. 3, pp. 1705–1721, 2010.

[75] S. R. Meshnick, "The mode of action of antimalarial endoperoxides," *Transactions of the Royal Society of Tropical Medicine and Hygiene*, vol. 88, pp. 31–32, 1994.

[76] B. Witkowski, J. Lelièvre, M.-L. Nicolau-Travers, X. Iriart, P. N. Soh, F. Bousejra-ElGarah, B. Meunier, A. Berry, and F. Benoit-Vical, "Evidence for the contribution of the hemozoin synthesis pathway of the murine plasmodium yoelii to the resistance to artemisinin-related drugs," *PLoS One*, vol. 7, no. 3, p. e32620, 2012.

[77] D. J. Creek, W. N. Charman, F. C. Chiu, R. J. Prankerd, Y. Dong, J. L. Vennerstrom, and S. A. Charman, "Relationship between antimalarial activity and heme alkylation for spiro-and dispiro-1, 2, 4-trioxolane antimalarials," *Antimicrobial agents and chemotherapy*, vol. 52, no. 4, pp. 1291–1296, 2008.

[78] S. R. Meshnick, R. K. Haynes, D. Monti, D. Taramelli, N. Basilico, S. Parapini, and P. Olliaro, "Artemisinin and heme," *Antimicrobial agents and chemotherapy*, vol. 47, no. 8, pp. 2712–2713, 2003.

[79] R. K. Haynes, D. Monti, D. Taramelli, N. Basilico, S. Parapini, and P. Olliaro, "Artemisinin antimalarials do not inhibit hemozoin formation," *Antimicrobial agents and chemotherapy*, vol. 47, no. 3, pp. 1175–1175, 2003.

[80] S. Krishna, C. J. Woodrow, H. M. Staines, R. K. Haynes, and O. Mercereau-Puijalon, "Re-evaluation of how artemisinins work in light of emerging evidence of in vitro resistance," *Trends in molecular medicine*, vol. 12, no. 5, pp. 200–205, 2006.

[81] S. R. Meshnick, "Artemisinin: mechanisms of action, resistance and toxicity," *International journal for parasitology*, vol. 32, no. 13, pp. 1655–1660, 2002.

[82] U. Eckstein-Ludwig, R. Webb, I. Van Goethem, J. East, *et al.*, "Artemisinins target the SERCA of Plasmodium falciparum," *Nature*, vol. 424, no. 6951, p. 957, 2003.

[83] D. Cardi, A. Pozza, B. Arnou, E. Marchal, J. D. Clausen, J. P. Andersen, S. Krishna, J. V. Møller, M. le Maire, and C. Jaxel, "Purified e255l mutant serca1a and purified pfatp6 are sensitive to serca-type inhibitors but insensitive to artemisinins," *Journal of Biological Chemistry*, vol. 285, no. 34, pp. 26406–26416, 2010.

[84] W. Li, W. Mo, D. Shen, L. Sun, J. Wang, S. Lu, J. M. Gitschier, and B. Zhou, "Yeast model uncovers dual roles of mitochondria in the action of artemisinin," *PLoS Genetics*, vol. 1, no. 3, p. e36, 2005.

[85] C. Sun, Y. Cao, P. Zhu, and B. Zhou, "A mitochondria-targeting artemisinin derivative with sharply increased antitumor but depressed anti-yeast and anti-malaria activities," *Scientific Reports*, vol. 7, 2017.

[86] C. Sun and B. Zhou, "The antimalarial drug artemisinin induces an additional, sod1-supressible anti-mitochondrial action in yeast," *Biochimica et Biophysica Acta (BBA)-Molecular Cell Research*, vol. 1864, no. 7, pp. 1285–1294, 2017.

[87] A. B. Vaidya, "Naphthoquinones: atovaquone, and other antimalarials targeting mitochondrial functions," in *Treatment and Prevention of Malaria*, pp. 127–139, Springer, 2011.

[88] H. Noedl, "Combination therapy in light of emerging artemisinin resistance," in *Treatment and Prevention of Malaria*, pp. 213–225, Springer, 2011.

[89] W. Peters, "Drug resistance in malaria—a perspective," *Transactions of the Royal Society of Tropical Medicine and Hygiene*, vol. 63, no. 1, pp. 25–40, 1969.

[90] N. White, "Preventing antimalarial drug resistance through combinations," *Drug resistance updates*, vol. 1, no. 1, pp. 3–9, 1998.

[91] P. Chen, G. Li, X. Guo, K. He, Y. Fu, L. Fu, and Y. Song, "The infectivity of gametocytes of Plasmodium falciparum from patients treated with artemisinin.," *Chinese medical journal*, vol. 107, no. 9, pp. 709–711, 1994.

[92] C. C. Campbell, "Malaria control—addressing challenges to ambitious goals," 2009.

[93] K. Ilett and K. Batty, "Artemisinin and its derivatives," in *Antimicrobial therapy and vaccines*, pp. 981–1002, ESun Technologies, 2005.

[94] I. M. Hastings, W. M. Watkins, and N. J. White, "The evolution of drug–resistant malaria: the role of drug elimination half–life," *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, vol. 357, no. 1420, pp. 505–519, 2002.

[95] K. Stepniewska and N. White, "Pharmacokinetic determinants of the window of selection for antimalarial drug resistance," *Antimicrobial agents and chemotherapy*, vol. 52, no. 5, pp. 1589–1596, 2008.

[96] A. Afonso, P. Hunt, S. Cheesman, A. C. Alves, C. Cunha, V. Do Rosário, and P. Cravo, "Malaria parasites can develop stable resistance to artemisinin but lack mutations in candidate genes atp6 (encoding the sarcoplasmic and endoplasmic reticulum $Ca^{2+}$ ATPase), tctp, mdr1, and cg10," *Antimicrobial agents and chemotherapy*, vol. 50, no. 2, pp. 480–489, 2006.

[97] L. K. Basco, J. Bickii, and P. Ringwald, "In vitro activity of lumefantrine (benflumetol) against clinical isolates of plasmodium falciparum in yaounde, cameroon," *Antimicrobial agents and chemotherapy*, vol. 42, no. 9, pp. 2347–2351, 1998.

[98] G. Holmgren, J. P. Gil, P. M. Ferreira, M. I. Veiga, C. O. Obonyo, and A. Björkman, "Amodiaquine resistant plasmodium falciparum malaria in vivo is associated with selection of pfcrt 76T and pfmdr1 86Y," *Infection, Genetics and Evolution*, vol. 6, no. 4, pp. 309–314, 2006.

[99] K. R. Tan, A. J. Magill, M. E. Parise, and P. M. Arguin, "Doxycycline for malaria chemoprophylaxis and treatment: report from the CDC expert meeting on malaria chemoprophylaxis," *The American journal of tropical medicine and hygiene*, vol. 84, no. 4, pp. 517–531, 2011.

[100] A. M. van Eijk and D. J. Terlouw, "Azithromycin for treating uncomplicated malaria," *Cochrane Database of Systematic Reviews*, no. 2, 2011. INFECTN.

[101] V. R. Solomon, S. K. Puri, K. Srivastava, and S. Katti, "Design and synthesis of new antimalarial agents from 4-aminoquinoline," *Bioorganic & medicinal chemistry*, vol. 13, no. 6, pp. 2157–2165, 2005.

[102] M. K. Gupta and Y. S. Prabhakar, "Topological descriptors in modeling the antimalarial activity of 4-(3', 5'-disubstituted anilino) quinolines," *Journal of chemical information and modeling*, vol. 46, no. 1, pp. 93–102, 2006.

[103] S. Deshpande, V. R. Solomon, S. B. Katti, and Y. S. Prabhakar, "Topological descriptors in modelling antimalarial activity: N 1-(7-chloro-4-quinolyl)-1, 4-bis

(3-aminopropyl) piperazine as prototype," *Journal of enzyme inhibition and medicinal chemistry*, vol. 24, no. 1, pp. 94–104, 2009.

[104] F. Luan, X. Xu, M. Natalia Dias Soeiro Cordeiro, H. Liu, and X. Zhang, "QSAR modeling for the antimalarial activity of 1, 4-naphthoquinonyl derivatives as potential antimalarial agents," *Current computer-aided drug design*, vol. 9, no. 1, pp. 95–107, 2013.

[105] P. K. Ojha and K. Roy, "Comparative QSARs for antimalarial endochins: importance of descriptor-thinning and noise reduction prior to feature selection," *Chemometrics and Intelligent Laboratory Systems*, vol. 109, no. 2, pp. 146–161, 2011.

[106] D. Hecht, M. Cheung, and G. B. Fogel, "QSAR using evolved neural networks for the inhibition of mutant pfDHFR by pyrimethamine derivatives," *Biosystems*, vol. 92, no. 1, pp. 10–15, 2008.

[107] A. Beheshti, E. Pourbasheer, M. Nekoei, and S. Vahdani, "QSAR modeling of antimalarial activity of urea derivatives using genetic algorithm–multiple linear regressions," *Journal of Saudi Chemical Society*, vol. 20, no. 3, pp. 282–290, 2016.

[108] P. A. d. S. Autreto and F. C. Lavarda, "Febrifugine derivative antimalarial activity: quantum mechanical predictors," *Revista do Instituto de Medicina Tropical de São Paulo*, vol. 50, no. 1, pp. 21–24, 2008.

[109] L. J. de Campos and E. B. de Melo, "Modeling structure–activity relationships of prodiginines with antimalarial activity using GA/MLR and OPS/PLS," *Journal of Molecular Graphics and Modelling*, vol. 54, pp. 19–31, 2014.

[110] N. Adhikari, A. K. Halder, C. Mondal, and T. Jha, "Exploring structural requirements of aurone derivatives as antimalarials by validated DFT-based QSAR, HQSAR, and COMFA–COMSIA approach," *Medicinal Chemistry Research*, vol. 22, no. 12, pp. 6029–6045, 2013.

[111] A. R. Katritzky, O. V. Kulshyn, I. Stoyanova-Slavova, D. A. Dobchev, M. Kuanar, D. C. Fara, and M. Karelson, "Antimalarial activity: a QSAR modeling using CODESSA PRO software," *Bioorganic & medicinal chemistry*, vol. 14, no. 7, pp. 2333–2357, 2006.

[112] C. Xue, S. Cui, M. Liu, Z. Hu, and B. Fan, "3D QSAR studies on antimalarial alkoxylated and hydroxylated chalcones by CoMFA and CoMSIA," *European journal of medicinal chemistry*, vol. 39, no. 9, pp. 745–753, 2004.

[113] A. K. Bhattacharjee, D. E. Kyle, J. L. Vennerstrom, and W. K. Milhous, "A 3D QSAR pharmacophore model and quantum chemical structure- activity analysis of chloroquine (CQ)-resistance reversal," *Journal of chemical information and computer sciences*, vol. 42, no. 5, pp. 1212–1220, 2002.

[114] G. Bringmann and C. Rummey, "3D QSAR investigations on antimalarial naphthylisoquinoline alkaloids by comparative molecular similarity indices analysis (CoMSIA), based on different alignment approaches," *Journal of chemical information and computer sciences*, vol. 43, no. 1, pp. 304–316, 2003.

[115] I. Garcia, Y. Fall, and G. Gomez, "QSAR, docking, and CoMFA studies of GSK3 inhibitors," *Current pharmaceutical design*, vol. 16, no. 24, pp. 2666–2675, 2010.

[116] F. Cheng, J. Shen, X. Luo, W. Zhu, J. Gu, R. Ji, H. Jiang, and K. Chen, "Molecular docking and 3D-QSAR studies on the possible antimalarial mechanism of artemisinin analogues," *Bioorganic & medicinal chemistry*, vol. 10, no. 9, pp. 2883–2891, 2002.

[117] C. Portela, C. M. Afonso, M. M. Pinto, and M. J. Ramos, "Receptor–drug association studies in the inhibition of the hematin aggregation process of malaria," *FEBS letters*, vol. 547, no. 1-3, pp. 217–222, 2003.

[118] L. Zhang, D. Fourches, A. Sedykh, H. Zhu, A. Golbraikh, S. Ekins, J. Clark, M. C. Connelly, M. Sigal, D. Hodges, *et al.*, "Discovery of novel antimalarial compounds enabled by qsar-based virtual screening," *Journal of chemical information and modeling*, vol. 53, no. 2, pp. 475–492, 2013.

[119] F.-J. Gamo, L. M. Sanz, J. Vidal, C. de Cozar, E. Alvarez, J.-L. Lavandera, D. E. Vanderwall, D. V. Green, V. Kumar, S. Hasan, *et al.*, "Thousands of chemical starting points for antimalarial lead identification," *Nature*, vol. 465, no. 7296, p. 305, 2010.

[120] P. Ludin, B. Woodcroft, S. A. Ralph, and P. Mäser, "In silico prediction of antimalarial drug target candidates," *International journal for parasitology: drugs and drug resistance*, vol. 2, pp. 191–199, 2012.

[121] A. Spitzmüller and J. Mestres, "Prediction of the P. falciparum target space relevant to malaria drug discovery," *PLoS computational biology*, vol. 9, no. 10, p. e1003257, 2013.

[122] M. Wawer and J. Bajorath, "Extracting SAR information from a large collection of anti-malarial screening hits by NSG-SPT analysis," *ACS medicinal chemistry letters*, vol. 2, no. 3, pp. 201–206, 2011.

[123] T. Spangenberg, J. N. Burrows, P. Kowalczyk, S. McDonald, T. N. Wells, and P. Willis, "The open access malaria box: a drug discovery catalyst for neglected diseases," *PloS one*, vol. 8, no. 6, p. e62906, 2013.

[124] A. Cherkasov, E. N. Muratov, D. Fourches, A. Varnek, I. I. Baskin, M. Cronin, J. Dearden, P. Gramatica, Y. C. Martin, R. Todeschini, V. Consonni, V. E. Kuz'min, R. Cramer, R. Benigni, C. Yang, J. Rathman, L. Terfloth, J. Gasteiger, A. Richard, and A. Tropsha, "QSAR modeling: Where have you been? where are you going to?," *Journal of Medicinal Chemistry*, vol. 57, no. 12, pp. 4977–5010, 2014.

[125] R. Todeschini and V. Consonni, *Handbook of molecular descriptors*, vol. 11. John Wiley & Sons, 2008.

[126] A. Varnek, D. Fourches, F. Hoonakker, and V. Soloviev, "Substructural fragments: an universal language to encode reactions, molecular and supramolecular structures," *Journal of Computer-Aided Molecular Design*, vol. 19, no. 9-10, pp. 693–703, 2005.

[127] F. Ruggiu, G. Marcou, A. Varnek, and D. Horvath, "ISIDA property-labelled fragment descriptors," *Molecular Informatics*, vol. 29, no. 12, pp. 855–868, 2010.

[128] M. Elhabiri, P. Sidorov, E. Cesar-Rodo, G. Marcou, D. A. Lanfranchi, E. Davioud-Charvet, D. Horvath, and A. Varnek, "Electrochemical properties of substituted 2-methyl-1, 4-naphthoquinones: Redox behavior predictions," *Chemistry-A European Journal*, vol. 21, no. 8, pp. 3415–3424, 2015.

[129] J. B. Mitchell, "Machine learning methods in chemoinformatics," *Wiley Interdisciplinary Reviews: Computational Molecular Science*, vol. 4, no. 5, pp. 468–481, 2014.

[130] T. M. Mitchell, "Artificial neural networks," *Machine learning*, vol. 45, pp. 81–127, 1997.

[131] G. Grégoire, "Multiple linear regression," *European Astronomical Society Publications Series*, vol. 66, pp. 45–72, 2014.

[132] V. Svetnik, A. Liaw, C. Tong, J. C. Culberson, R. P. Sheridan, and B. P. Feuston, "Random forest: a classification and regression tool for compound classification and QSAR modeling," *Journal of chemical information and computer sciences*, vol. 43, no. 6, pp. 1947–1958, 2003.

[133] C. Cortes and V. Vapnik, "Support vector machine," *Machine learning*, vol. 20, no. 3, pp. 273–297, 1995.

[134] A. K. Jain, M. N. Murty, and P. J. Flynn, "Data clustering: a review," *ACM computing surveys (CSUR)*, vol. 31, no. 3, pp. 264–323, 1999.

[135] A. K. Jain, "Data clustering: 50 years beyond k-means," *Pattern recognition letters*, vol. 31, no. 8, pp. 651–666, 2010.

[136] U. M. Fayyad, A. Wierse, and G. G. Grinstein, *Information visualization in data mining and knowledge discovery*. Morgan Kaufmann, 2002.

[137] F. Korn, B.-U. Pagel, and C. Faloutsos, "On the "dimensionality curse" and the "self-similarity blessing"," *IEEE Transactions on Knowledge and Data Engineering*, vol. 13, no. 1, pp. 96–111, 2001.

[138] S. Wold, K. Esbensen, and P. Geladi, "Principal component analysis," *Chemometrics and intelligent laboratory systems*, vol. 2, no. 1-3, pp. 37–52, 1987.

[139] A. J. Izenman, "Linear discriminant analysis," in *Modern multivariate statistical techniques*, pp. 237–280, Springer, 2013.

[140] M. L. Davison, "Multidimensional scaling," *New York*, 1983.

[141] S. T. Roweis and L. K. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *science*, vol. 290, no. 5500, pp. 2323–2326, 2000.

[142] T. Kohonen, "The self-organizing map," *Neurocomputing*, vol. 21, no. 1, pp. 1–6, 1998.

[143] C. M. Bishop, M. Svensén, and C. K. Williams, "GTM: The generative topographic mapping," *Neural computation*, vol. 10, no. 1, pp. 215–234, 1998.

[144] A. R. Leach and V. J. Gillet, *An Introduction to Chemoinformatics*. Springer Publishing Company, Incorporated, 2007.

[145] M. W. Browne, "Cross-validation methods," *Journal of mathematical psychology*, vol. 44, no. 1, pp. 108–132, 2000.

[146] M. Svensén, *GTM: The Generative Topographic Mapping*. PhD thesis, University of Aston in Birmingham, 1998.

[147] H. A. Gaspar, P. Sidorov, D. Horvath, I. I. Baskin, G. Marcou, and A. Varnek, "Generative Topographic Mapping approach to chemical space analysis," in *Frontiers in Molecular Design and Chemical Information Science - Herman Skolnik Award Symposium 2015: Jürgen Bajorath*, pp. 211–241, 2016.

[148] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society, Series B*, vol. 39, no. 1, pp. 1–38, 1977.

[149] H. A. Gaspar, I. I. Baskin, G. Marcou, D. Horvath, and A. Varnek, "Chemical data visualization and analysis with incremental generative topographic mapping: big data challenge," *Journal of chemical information and modeling*, vol. 55, no. 1, pp. 84–94, 2014.

[150] H. Gaspar, I. Baskin, G. Marcou, D. Horvath, and A. Varnek, "GTM-based QSAR models and their applicability domains," *Molecular informatics*, vol. 34, no. 6-7, pp. 348–356, 2015.

[151] H. A. Gaspar, G. Marcou, D. Horvath, A. Arault, S. Lozano, P. Vayer, and A. Varnek, "Generative Topographic Mapping-based classification models and their applicability domain: Application to the biopharmaceutics drug disposition classification system (BDDCS)," *Journal of Chemical Information and Modeling*, vol. 53, no. 12, pp. 3318–3325, 2013.

[152] N. Kireeva, I. I. Baskin, H. A. Gaspar, D. Horvath, G. Marcou, and A. Varnek, "Generative Topographic Mapping (GTM): Universal tool for data visualization, structure-activity modeling and dataset comparison," *Molecular Informatics*, vol. 31, no. 3-4, pp. 301–312, 2012.

[153] I. M. Grimmenstein and W. Urfer, "Analyzing protein data with the generative topographic mapping approach," in *Innovations in Classification, Data Science, and Information Systems* (D. Baier and K.-D. Wernecke, eds.), Studies in Classification, Data Analysis, and Knowledge Organization, pp. 585–592, Springer Berlin Heidelberg, 2005.

[154] D. Horvath, I. Baskin, G. Marcou, and A. Varnek, "Generative Topographic Mapping of conformational space," *Molecular Informatics*, 2017.

[155] K. Klimenko, G. Marcou, D. Horvath, and A. Varnek, "Chemical space mapping and structure-activity analysis of the ChEMBl antiviral compound set," *Journal of chemical information and modeling*, vol. 56, no. 8, pp. 1438–1454, 2016.

[156] T. I. Netzeva, A. P. Worth, T. Aldenberg, R. Benigni, M. T. Cronin, P. Gramatica, J. S. Jaworska, S. Kahn, G. Klopman, C. A. Marchant, *et al.*, "Current status of methods for defining the applicability domain of (quantitative) structure-activity relationships," *ATLA*, vol. 33, pp. 155–173, 2005.

[157] T. I. Oprea and J. Gottfries, "Chemography: the art of navigating in chemical space," *Journal of combinatorial chemistry*, vol. 3, no. 2, pp. 157–166, 2001.

[158] S. Kayastha, D. Horvath, E. Gilberg, M. Gütschow, J. Bajorath, and A. Varnek, "Privileged Structural Motif detection and analysis using Generative

Topographic Maps," *Journal of Chemical Information and Modeling*, vol. 57, no. 5, pp. 1218–1232, 2017.

[159] A. Tropsha, "Best practices for QSAR model development, validation, and exploitation," *Molecular informatics*, vol. 29, no. 6-7, pp. 476–488, 2010.

[160] ChEMBL database. https://www.ebi.ac.uk/chembl/, 2017.

[161] B. Viira, T. Gendron, D. A. Lanfranchi, S. Cojean, D. Horvath, G. Marcou, A. Varnek, L. Maes, U. Maran, P. M. Loiseau, *et al.*, "In silico mining for antimalarial structure-activity knowledge and discovery of novel antimalarial curcuminoids," *Molecules*, vol. 21, no. 7, p. 853, 2016.

[162] W. A. Guiguemde, A. A. Shelat, D. Bouck, S. Duffy, G. J. Crowther, P. H. Davis, D. C. Smithson, M. Connelly, J. Clark, F. Zhu, M. B. Jiménez-Díaz, M. S. Martinez, E. B. Wilson, A. K. Tripathi, J. Gut, E. R. Sharlow, I. Bathurst, F. E. Mazouni, J. W. Fowble, I. Forquer, P. L. McGinley, S. Castro, I. Angulo-Barturen, S. Ferrer, P. J. Rosenthal, J. L. DeRisi, D. J. Sullivan, J. S. Lazo, D. S. Roos, M. K. Riscoe, M. A. Phillips, P. K. Rathod, W. C. Van Voorhis, V. M. Avery, and R. K. Guy, "Chemical genetics of Plasmodium falciparum," *Nature*, vol. 465, pp. 311–315, May 2010.

[163] T. Joët and S. Krishna, "The hexose transporter of Plasmodium falciparum is a worthy drug target," *Acta tropica*, vol. 89, no. 3, pp. 371–374, 2004.

[164] L. Luzzatto, E. A. Usanga, and S. Reddy, "Glucose-6-phosphate dehydrogenase deficient red cells: resistance to infection by malarial parasites," *Science*, vol. 164, no. 3881, pp. 839–842, 1969.

[165] M. Nankya-Kitaka, G. Curley, C. Gavigan, A. Bell, and J. Dalton, "Plasmodium chabaudi chabaudi and P. falciparum: inhibition of aminopeptidase and parasite growth by bestatin and nitrobestatin," *Parasitology research*, vol. 84, no. 7, pp. 552–558, 1998.

[166] ChemAxon Standardizer. https://docs.chemaxon.com/display/docs/Standardizer+User%27s+Guide.

[167] ChemAxon Tautomer plugin. https://docs.chemaxon.com/display/docs/Tautomer+Generation+Plugin.

[168] ISIDA Fragmentor. http://infochim.u-strasbg.fr/recherche/Download/Fragmentor/Fragmentor2015_Manual.pdf, 2015.

[169] A. Schuffenhauer, P. Ertl, S. Roggo, S. Wetzel, M. A. Koch, and H. Waldmann, "The scaffold tree – visualization of the scaffold universe by hierarchical scaffold

classification," *Journal of chemical information and modeling*, vol. 47, no. 1, pp. 47–58, 2007.

[170] D. G. Lloyd, G. Golfis, A. J. Knox, D. Fayne, M. J. Meegan, and T. I. Oprea, "Oncology exploration: charting cancer medicinal chemistry space," *Drug discovery today*, vol. 11, no. 3, pp. 149–159, 2006.

[171] S. Matero, M. Lahtela-Kakkonen, O. Korhonen, J. Ketolainen, R. Lappalainen, and A. Poso, "Chemical space of orally active compounds," *Chemometrics and intelligent laboratory systems*, vol. 84, no. 1, pp. 134–141, 2006.

[172] T. Öberg and M. S. Iqbal, "The chemical and environmental property space of REACH chemicals," *Chemosphere*, vol. 87, no. 8, pp. 975–981, 2012.

[173] L. M. Kauvar, H. O. Villar, J. R. Sportsman, D. L. Higgins, and D. E. Schmidt, "Protein affinity map of chemical space," *Journal of Chromatography B: Biomedical Sciences and Applications*, vol. 715, no. 1, pp. 93–102, 1998.

[174] D. Horvath, M. Lisurek, B. Rupp, R. Kühne, E. Specker, J. von Kries, D. Rognan, C. D. Andersson, F. Almqvist, M. Elofsson, *et al.*, "Design of a general-purpose european compound screening library for EU-OPENSCREEN," *ChemMedChem*, vol. 9, no. 10, pp. 2309–2326, 2014.

[175] C. Abad-Zapatero, O. Perišić, J. Wass, A. P. Bento, J. Overington, B. Al-Lazikani, and M. E. Johnson, "Ligand efficiency indices for an effective mapping of chemico-biological space: the concept of an atlas-like representation," *Drug discovery today*, vol. 15, no. 19, pp. 804–811, 2010.

[176] G. V. Paolini, R. H. Shapland, W. P. van Hoorn, J. S. Mason, and A. L. Hopkins, "Global mapping of pharmacological space," *Nat. Biotechnol.*, vol. 24, pp. 805–815, Jul 2006.

[177] L. Ruddigkeit, R. Van Deursen, L. C. Blum, and J.-L. Reymond, "Enumeration of 166 billion organic small molecules in the chemical universe database GDB-17," *Journal of chemical information and modeling*, vol. 52, no. 11, pp. 2864–2875, 2012.

[178] J.-L. Reymond, L. Ruddigkeit, L. Blum, and R. van Deursen, "The enumeration of chemical space," *Wiley Interdisciplinary Reviews: Computational Molecular Science*, vol. 2, no. 5, pp. 717–733, 2012.

[179] P. G. Polishchuk, T. I. Madzhidov, and A. Varnek, "Estimation of the size of drug-like chemical space based on GDB-17 data," *Journal of computer-aided molecular design*, vol. 27, no. 8, pp. 675–679, 2013.

[180] F. Bonachera, G. Marcou, N. Kireeva, A. Varnek, and D. Horvath, "Using self-organizing maps to accelerate similarity search," *Bioorganic & medicinal chemistry*, vol. 20, no. 18, pp. 5396–5409, 2012.

[181] D. Horvath, C. Koch, G. Schneider, G. Marcou, and A. Varnek, "Local neighborhood behavior in a combinatorial library context," *Journal of computer-aided molecular design*, vol. 25, no. 3, pp. 237–252, 2011.

[182] D. Horvath and C. Jeandenans, "Neighborhood behavior of in silico structural spaces with respect to in vitro activity spaces - a benchmark for neighborhood behavior assessment of different in silico similarity metrics," *Journal of chemical information and computer sciences*, vol. 43, no. 2, pp. 691–698, 2003.

[183] N. Huang, B. K. Shoichet, and J. J. Irwin, "Benchmarking sets for molecular docking," *Journal of medicinal chemistry*, vol. 49, no. 23, pp. 6789–6801, 2006.

[184] H. A. Gaspar, *Cartographie de l'espace chimique*. PhD thesis, Université de Strasbourg, 2015.

[185] L. Davis, "Handbook of genetic algorithms," 1991.

[186] D. E. Goldberg and M. P. Samtani, "Engineering optimization via genetic algorithm," in *Electronic computation*, pp. 471–482, ASCE, 1986.

[187] E. J. Gardiner, P. Willett, and P. J. Artymiuk, "Protein docking using a genetic algorithm," *Proteins: Structure, Function, and Bioinformatics*, vol. 44, no. 1, pp. 44–56, 2001.

[188] B. K. Lavine, C. Davidson, and A. J. Moores, "Innovative genetic algorithms for chemoinformatics," *Chemometrics and Intelligent Laboratory Systems*, vol. 60, no. 1, pp. 161–171, 2002.

[189] P. Vas, *Artificial-intelligence-based electrical machines and drives: application of fuzzy, neural, fuzzy-neural, and genetic-algorithm-based techniques*, vol. 45. Oxford university press, 1999.

[190] D. Horvath, J. Brown, G. Marcou, and A. Varnek, "An evolutionary optimizer of libsvm models," *Challenges*, vol. 5, no. 2, pp. 450–472, 2014.

[191] J. J. Irwin and B. K. Shoichet, "ZINC – a free database of commercially available compounds for virtual screening," *Journal of chemical information and modeling*, vol. 45, no. 1, pp. 177–182, 2005.

[192] T. R. Gimadiev, T. I. Madzhidov, G. Marcou, and A. Varnek, "Generative Topographic Mapping approach to modeling and chemical space visualization of human intestinal transporters," *BioNanoScience*, vol. 6, no. 4, pp. 464–472, 2016.

[193] J. L. Wang, K. Aston, D. Limburg, C. Ludwig, A. E. Hallinan, F. Koszyk, B. Hamper, D. Brown, M. Graneto, J. Talley, *et al.*, "The novel benzopyran class of selective cyclooxygenase-2 inhibitors. part III: The three microdose candidates," *Bioorganic & medicinal chemistry letters*, vol. 20, no. 23, pp. 7164–7168, 2010.

[194] J. M. Janusz, P. A. Young, J. M. Ridgeway, M. W. Scherz, K. Enzweiler, L. I. Wu, L. Gan, J. Chen, D. E. Kellstein, S. A. Green, *et al.*, "New cyclooxygenase-2/5-lipoxygenase inhibitors. 3. 7-tert-butyl-2, 3-dihydro-3, 3-dimethylbenzofuran derivatives as gastrointestinal safe antiinflammatory and analgesic agents: variations at the 5 position," *Journal of medicinal chemistry*, vol. 41, no. 18, pp. 3515–3529, 1998.

[195] K. A. Jacobson, P. J. Van Galen, and M. Williams, "Adenosine receptors: pharmacology, structure-activity relationships, and therapeutic potential," *Journal of medicinal chemistry*, vol. 35, no. 3, pp. 407–422, 1992.

[196] S.-A. Poulsen and R. J. Quinn, "Adenosine receptors: new opportunities for future drugs," *Bioorganic & medicinal chemistry*, vol. 6, no. 6, pp. 619–641, 1998.

[197] S. P. Davies, H. Reddy, M. Caivano, and P. Cohen, "Specificity and mechanism of action of some commonly used protein kinase inhibitors," *Biochemical Journal*, vol. 351, no. 1, pp. 95–105, 2000.

[198] J. Bain, L. Plater, M. Elliott, N. Shpiro, C. J. Hastie, H. Mclauchlan, I. Klevernic, J. S. C. Arthur, D. R. Alessi, and P. Cohen, "The selectivity of protein kinase inhibitors: a further update," *Biochemical Journal*, vol. 408, no. 3, pp. 297–315, 2007.

[199] J. B. Baell and G. A. Holloway, "New substructure filters for removal of pan assay interference compounds (PAINS) from screening libraries and for their exclusion in bioassays," *Journal of medicinal chemistry*, vol. 53, no. 7, pp. 2719–2740, 2010.

[200] P. Sidorov, I. Desta, M. Chessé, D. Horvath, G. Marcou, A. Varnek, E. Davioud-Charvet, and M. Elhabiri, "Redox polypharmacology as an emerging strategy to combat malarial parasites," *ChemMedChem*, vol. 11, no. 12, pp. 1339–1351, 2016.

[201] T. Müller, L. Johann, B. Jannack, M. Brückner, D. A. Lanfranchi, H. Bauer, C. Sanchez, V. Yardley, C. Deregnaucourt, J. Schrével, *et al.*, "Glutathione reductase-catalyzed cascade of redox reactions to bioactivate potent antimalarial 1, 4-naphthoquinones – a new strategy to combat malarial parasites," *Journal of the American Chemical Society*, vol. 133, no. 30, pp. 11557–11571, 2011.

[202] D. A. Lanfranchi, E. Cesar-Rodo, B. Bertrand, H.-H. Huang, L. Day, L. Johann, M. Elhabiri, K. Becker, D. L. Williams, and E. Davioud-Charvet, "Synthesis and biological evaluation of 1, 4-naphthoquinones and quinoline-5, 8-diones as antimalarial and schistosomicidal agents," *Organic & biomolecular chemistry*, vol. 10, no. 31, pp. 6375–6387, 2012.

[203] T. J. Egan, R. Hunter, C. H. Kaschula, H. M. Marques, A. Misplon, and J. Walden, "Structure- function relationships in aminoquinolines: effect of amino and chloro groups on quinoline- hematin complex formation, inhibition of $\beta$-hematin formation, and antiplasmodial activity," *Journal of medicinal chemistry*, vol. 43, no. 2, pp. 283–291, 2000.

[204] R. L. Krauth-Siegel, J. G. Müller, F. Lottspeich, and R. H. Schirmer, "Glutathione reductase and glutamate dehydrogenase of Plasmodium falciparum, the causative agent of tropical malaria," *The FEBS Journal*, vol. 235, no. 1-2, pp. 345–350, 1996.

[205] K. Reybier, T. H. Y. Nguyen, H. Ibrahim, P. Perio, A. Montrose, P.-L. Fabre, and F. Nepveu, "Electrochemical behavior of indolone-N-oxides: relationship to structure and antiplasmodial activity," *Bioelectrochemistry*, vol. 88, pp. 57–64, 2012.

[206] V. Alagarsamy, *Textbook of Medicinal Chemistry*, vol. 1. Elsevier Health Sciences, 2013.

[207] D. Harvey, *Modern analytical chemistry*, vol. 381. McGraw-Hill New York, 2000.

[208] T. Rodríguez-Fernández, V. Ugalde-Saldívar, I. González, L. Escobar, and J. García-Valdés, "Electrochemical strategy to scout 1, 4-naphthoquinones effect on voltage gated potassium channels," *Bioelectrochemistry*, vol. 86, pp. 1–8, 2012.

[209] K. Alizadeh and M. Shamsipur, "Calculation of the two-step reduction potentials of some quinones in acetonitrile," *Journal of Molecular Structure: THEOCHEM*, vol. 862, no. 1, pp. 39–43, 2008.

[210] J. L. Cape, M. K. Bowman, and D. M. Kramer, "Computation of the redox and protonation properties of quinones: towards the prediction of redox cycling natural products," *Phytochemistry*, vol. 67, no. 16, pp. 1781–1788, 2006.

[211] M. Namazian, P. Norouzi, and R. Ranjbar, "Prediction of electrode potentials of some quinone derivatives in acetonitrile," *Journal of Molecular Structure: THEOCHEM*, vol. 625, no. 1, pp. 235–241, 2003.

[212] M. Namazian and P. Norouzi, "Prediction of one-electron electrode potentials of some quinones in dimethylsulfoxide," *Journal of Electroanalytical Chemistry*, vol. 573, no. 1, pp. 49–53, 2004.

[213] M. R. Hadjmohammadi, K. Kamel, and P. Biparva, "Quantitative structure-reduction potential relationship study of some quinones in five solvents," *Journal of solution chemistry*, vol. 40, no. 2, pp. 224–230, 2011.

[214] A. Beheshti, P. Norouzi, and M. Ganjali, "A simple and robust model for predicting the reduction potential of quinones family; Electrophilicity Index effect," *Int J Electrochem Sci*, vol. 7, pp. 4811–4821, 2012.

[215] P. Thanikaivelan, V. Subramanian, J. R. Rao, and B. U. Nair, "Application of quantum chemical descriptor in quantitative structure activity and structure property relationship," *Chemical Physics Letters*, vol. 323, no. 1, pp. 59–70, 2000.

[216] J. Gasteiger and M. Marsili, "A new model for calculating atomic charges in molecules," *Tetrahedron Letters*, vol. 19, no. 34, pp. 3181–3184, 1978.

[217] ChemAxon Calculation of Partial Charge Distributions. `http://www.chemaxon.com/marvin/help/calculations/charge.html`. 2008.

[218] ChemAxon Polarizability plugin. `http://www.chemaxon.com/marvinarchive/4.0/marvin/chemaxon/marvin/help/calculator-plugins.html#polarizability`. 2008.

[219] ChemAxon Hückel Analysis plugin. `http://www.chemaxon.com/marvin/help/calculations/other.html#huckel`. 2008.

[220] D. Horvath, G. Marcou, and A. Varnek, "Predicting the predictability: a unified approach to the applicability domain problem of QSAR models," *Journal of chemical information and modeling*, vol. 49, no. 7, pp. 1762–1776, 2009.

[221] I. V. Tetko, I. Sushko, A. K. Pandey, H. Zhu, A. Tropsha, E. Papa, T. Oberg, R. Todeschini, D. Fourches, and A. Varnek, "Critical assessment of qsar models of environmental toxicity against tetrahymena pyriformis: focusing on applicability domain and overfitting by variable selection," *Journal of chemical information and modeling*, vol. 48, no. 9, pp. 1733–1746, 2008.

[222] A. Varnek, D. Fourches, V. Solov'ev, O. Klimchuk, A. Ouadi, and I. Billard, "Successful "in silico" design of new efficient uranyl binders," *Solvent Extraction and Ion Exchange*, vol. 25, no. 4, pp. 433–462, 2007.

[223] C. Agostinelli, "Robust stepwise regression," *Journal of Applied Statistics*, vol. 29, no. 6, pp. 825–840, 2002.

[224] M. Bostock, V. Ogievetsky, and J. Heer, "$D^3$: Data-Driven Documents," *IEEE transactions on visualization and computer graphics*, vol. 17, no. 12, pp. 2301–2309, 2011.

[225] OpenChemLib. `https://github.com/actelion/openchemlib`.

[226] D. Lasso plugin. `https://github.com/skokenes/D3-Lasso-Plugin`.

# Appendix

## MalariaDB composition

The MalariaDB contains information on 15462 molecules, and in the database following fields are considered:

- **CHEMBLID** is the key field that distinguishes the compounds. If a compound comes from our colleagues and thus doesn't have a ChEMBL ID, its lab code is taken as the key.
- **SMILES** is, obviously, the SMILES representation of the chemical structure as given in ChEMBL.
- **alt_chid**. If a compound is present in ChEMBL as a salt or other alternative forms that is not distinguished after the standardization, its other ChEMBL IDs are indicated here.
- **Mechanisms** are separated by their source, so three distinct fields are responsible for the annotation (mechanism or target annotations by TCAMS (GSK_mech), MalariaBox(MBox_mech), or ActivityDB (SJ_mech)).
- **Target family** follows the grouping strategy explained in Chapter 4.
- **Set** corresponds to the set in which the compound is found (TCAMS, MalariaBox, or any of seventeen subsets of ActivityDB). If it is present in several sets, all are indicated.
- **SJ_code** is the compound code from Guiguemde et al., Nature 465, 2010.
- **ExpInfo Assay**. If a compound from a database have an experimental validation of mechanism from ChEMBL, the ID of the assay in which it was validated is indicated.
- **ExpInfo comment**. The important information from the assay results is kept here.

- **Confidence**. The confidence level of the annotation: experimental, by similarity, orthology, or enriched human target.

## ActivityDB subsets

The full information on experimental conditions for each of ActivityDB subsets is given in the table below. Where the set consists purely of ChEMBL data, its ChEMBL ID is given as the subset name. For details on the experimental parameters taken into account, refer to **Viira et al. Molecules 2016, 21, 853; doi:10.3390/molecules21070853**.

| Subset ID | Measured Property (Endpoint) | Plasmodium Strain | Drug Exposure Time | Hematocrit % | Parasitic Stage | Assay | Set Size |
|---|---|---|---|---|---|---|---|
| PS1 | pIC50 | 3D7 | 48 h | 5.0 | async | 3H-hyp | 65 |
| PS2 | pIC50 | K1 | 72 h | 2.5 | async | 3H-hyp | 126 |
| PS3 | pIC50 | Dd2 | 48 h | 1.5 | sync | 3H-hyp | 66 |
| PS4 | pIC50 | Dd2 | 48 h | 2.0 | sync | 3H-hyp | 70 |
| PS5 | pIC50 | K1 | 48 h | 2.5 | async | 3H-hyp | 125 |
| PS6 | pIC50 | 3D7 | 48 h | 2.0 | sync | 3H-hyp | 120 |
| PS7 | pIC50 | 3D7 | 48 h | 2.5 | async | 3H-hyp | 94 |
| PS8 | pIC50 | Dd2 | 72 h | 2.0 | async | SYBRg | 143 |
| PS9 | pIC50 | K1 | 72 h | 1.25 | async | 3H-hyp | 161 |
| PS10 | pIC50 | K1 | 48 h | 1.5 | aync | 3H-hyp | 67 |
| CHEMBL730080 | pEC50 | K1 | 72 h | 2.0 | sync | SYBRg | 989 |
| CHEMBL896244 | pED50 | 3D7 | 72 h | 0.5 | async | 3H-hyp | 230 |
| CHEMBL896245 | pED50 | K1 | 72 h | 0.5 | async | 3H-hyp | 201 |
| CHEMBL1038869 | pEC50 | SB-A6 | 72 h | 2.0 | sync | SYBRg | 163 |
| CHEMBL1038870 | pEC50 | D10 | 72 h | 2.0 | sync | SYBRg | 160 |
| CHEMBL730081 | pEC50 | 3D7 | 72 h | 2.0 | sync | SYBRg | 168 |
| CHEMBL730641 | pEC50 | K1 | 72 h | 2.0 | sync | SYBRg | 162 |