

ÉCOLE DOCTORALE DES SCIENCES CHIMIQUES
UMR 7140

THÈSE présentée par :
Kyrylo KLIMENKO
soutenue le: **14 mars 2017**

pour obtenir le grade de : **Docteur de l'université de Strasbourg**
Discipline/S spécialité : Chimie/Chémoinformatique

Computer-aided drug design of broad-spectrum antiviral compounds

THÈSE dirigée par :

M. VARNEK Alexandre	Professeur, Université de Strasbourg
M. KUZMIN Victor	Professeur, Institut de Chimie Physique, Odessa, Ukraine

RAPPORTEURS :

M. ERTL Peter	Docteur, HDR, société Novartis
M. TABOUREAU Olivier	Professeur, Université Paris Diderot

AUTRES MEMBRES DU JURY :

Mme. KELLENBERGER Esther	Professeur, Université de Strasbourg
---------------------------------	--------------------------------------

Mr Kyrylo Klimenko was a member of the European Doctoral College of the University of Strasbourg during the preparation of his PhD thesis (2013-2016), class Dmitri Mendeleev. In addition to his mainstream research topic, he has attended lectures and seminars dedicated to general European policies and values presented by international experts. This PhD research project was carried out in scope of collaboration between the University of Strasbourg, France and A.V. Bogatsky Physico-Chemical Institute, Ukraine.

Acknowledgements

First of all, I would like to express my gratitude to my academic advisors, Prof Alexandre Varnek and Prof Victor Kuz`min for their brilliant guidance in completing the academic challenges I have faced. I am also thankful to the French Embassy in Ukraine for providing financial means to the project's completion.

I would like to thank all the members of the jury, Prof Olivier Taboureau, Dr Peter Ertl and Prof Esther Kellenberger for accepting to judge and review my work.

I would also like to express my gratitude to my senior colleagues Dr. Dragos Horvath, Dr Gilles Marcou, Dr Pavel Polishchuk, as well as, former and current PhD students Dr. Helena Gaspar, Dr. Fiorela Ruggiu, Grace Delouis, Pavel Sidorov, Timur Gimadiev and Julien Denos for their advice and support.

I am thankful to my collaborators from Physico-chemical Institute, Ukraine and Institute of Chemical Biology and Fundamental Medicine Russia who provided experimental evidence to my theoretical findings.

Special thanks to Dr. Olga Klimchuk and Soumia Hnini for their help on accommodation and administration tasks.

I am deeply and forever indebted to my parents and friends for their love and support.

ABSTRACT

Virtual screening and cartography of chemical space approaches have been used for design of broad-spectrum antivirals acting as nucleic acids intercalators. The 1st part of thesis reports QSPR model for aqueous solubility of organic molecules within the wide temperature range. This model was later used for solubility assessment of antiviral compounds. In the second part of work, structural filters, QSAR and pharmacophore models were developed then used to screen a database containing some 3.2 M compounds. This resulted in 55 hits which were synthesized and experimentally tested. Two lead compounds displayed high activity against Vaccinia virus and low toxicity. In the 3d part of the thesis, Generative Topographic Mapping (GTM) approach was used to build 2D maps of chemical space of antiviral compounds. Experimental data on antiviral compounds were extracted from ChEMBL database, curated and annotated by major virus Genus. Selected dataset was used to build maps on which all other ChEMBL compounds were projected. Analysis of the maps revealed structural motifs characterizing particular types of antivirals.

DECLARATION

- This is my own work
- Information sources were acknowledged and fully referenced.

Kyrylo Klimenko
March, 2017

Abbreviations

AD : Applicability domain

BA: Balanced accuracy

BVDV : Bovine viral diarrhoea virus

CMV : Cytomegalovirus

FN : False positives

FP : False positives

GFP : Green fluorescent protein

GTM : Generative topographic mapping

HBV : Hepatitis B virus

HCV : Hepatitis C virus

HIV : Human immunodeficiency virus

HSV : Herpes simplex virus

ISIDA-SMF : In silico design and data analysis substructure molecular fragments

MRV : Mammalian orthoreovirus

OOB : Out-of-bag

PASS : Prediction of activity spectra for substances

PFU : Plaque forming unit

PSM : Privileged structural motifs

PRP : Privileged responsibility patterns

QSA(P)R : Quantitative structure-activity(property) relationship

RF : Random forest

RMSE : Root-mean-square-error

SiRMS : Simplex representation of molecular structure

TP : True positives

TN : True negatives

VSV : Vesicular stomatitis virus

XV : Cross-validation

Contents

Résumé en français.....	15
INTRODUCTION.....	25
PART 1 REVIEW ON VIRUS PROBLEMATICS	27
1.1 Overview of the virus structure and reproduction	27
1.2 Current antiviral treatment strategies.....	32
1.3 Earlier studies on computer-aided design of antiviral drugs.....	35
PART 2 COMPUTATIONAL TECHNIQUES USED IN THIS STUDY	39
2.1 (Q)SA(P)R approach - The (Quantitative) Structure-Activity (Property) Relationship	39
2.1.1 SiRMS (Simplex representation of molecular structure)	39
2.1.2 ISIDA substructure molecular fragments	40
2.1.3 Random Forest.....	42
2.1.4 Generative Topographic Method.....	43
2.1.5 Statistics used for QSAR models performance assessment	48
2.1.6 Pharmacophore modeling	50
2.1.7 Third-party predictive models used in Virtual Screening	53
2.2 Databases.....	53
2.3 Data curation tool.....	54
2.4 Scaffold analysis tool	56
PART 3 QSPR MODEL FOR AQUEOUS SOLUBILITY PREDICTION.....	61
3.1 Dataset preparation.....	61
3.2 Model development.....	63
3.2.1 Determination of solubility-temperature equation.....	63
3.2.2 QSPR model for <i>k_j</i> prediction.....	65
3.2.3 QSPR solubility model development	65
3.3 Model validation on external test set	65
3.4 Conclusions	67
PART 4 COMPUTER-AIDED DESIGN OF BROAD-SPECTRUM ANTIVIRALS	69
4.1 Data preparation	69
4.2 Modeling.....	71
4.2.1 Filters	71
4.2.2 Pharmacopore models.....	72

4.2.3 Classification SAR models	75
4.3 Virtual screening	75
4.4 Conclusions	80
PART 5 VISUALIZATION AND ANALYSIS OF CHEMICAL SPACE OF ANTIVIRAL COMPOUNDS	81
5.1 Data curation	82
5.2 Model development	89
5.3 Chemical space analysis	93
5.3.1 Visualization of the chemical space	93
5.3.2. Analysis of the privileged patterns	99
5.4 Activity prediction of antiviral CADD compounds	106
5.5 Conclusions	109
GENERAL CONCLUSIONS.....	113
REFERENCES.....	115
APPENDICES	125
APPENDIX A Supplementary material to QSPR modeling of aqueous solubility	126
APPENDIX B Supplementary material to computer-aided design of new antiviral compounds	137
APPENDIX C Published and submitted articles	163

Résumé en français

Conception assistée par ordinateur de composés antiviraux à large spectre

INTRODUCTION

Les infections virales sont à l'origine de nombreuses maladies dangereuses. Les médicaments antiviraux existants affectent essentiellement des protéines virales spécifiques inhibant la reproduction de virus spécifiques. Ces pratiques ne permettent pas de contourner les résistances développées par les virus, ni ne permettent de traiter simultanément plusieurs infections. Une stratégie de développement de composés antiviraux moins fréquente est de rechercher des molécules ayant un large spectre d'activité antivirale. Les principaux groupes d'antiviraux à large spectre incluent les analogues de nucléotides (p. ex. acyclovir) et des petites molécules inductrices d'interférons (p. ex. tilorone). Toutefois, les analogues de nucléotides sont surtout actifs sur des virus à ADN et des rétrovirus et, en raison de leur faible biodisponibilité orale et de leur toxicité significative, ont un intérêt limité pour le traitement de maladies chroniques pour lesquelles des thérapies par voie orale sont particulièrement recherchées. Les petites molécules inductrices d'interféron peuvent montrer une activité à la fois contre des virus à ARN et à ADN mais ils manquent d'efficacité pour l'élimination complète d'une infection virale.

Les intercalants d'acides nucléiques constituent une classe mésestimée d'antivirus à large spectre. L'intercalation change la conformation des acides nucléiques viraux conduisant à les rendre impropres à jouer leur rôle biologique et ainsi, empêchant la reproduction des virus à acides nucléiques à double brin. Cette propriété rend les intercalants d'acides nucléiques particulièrement attractifs pour le développement de nouveaux médicaments antiviraux.

L'objectif de ce travail est le développement assisté par ordinateur de nouveaux intercalants possédant un large spectre d'activité antiviral. Nous avons utilisé de nombreuses approches chémoinformatiques (filtres, QSAR et pharmacophores) pour

construire des modèles prédictifs utiles pour le criblage virtuel d'une chimiothèque de plus 3M de composés. Le criblage a débouché sur une sélection de 55 touches qui, tout d'abord, ont été synthétisés à l'Institut de Physico-Chimie A. V. Bogatski (PCI) à Odessa, Ukraine, puis testés expérimentalement à l'Institut de Biochimie et Médecine Fondamentale (ICBFM) à Novosibirsk, Russie. Deux molécules appartenant à la famille des indolequinaxolines ont été identifiées comme des intercalants d'ADN actifs contre le *Vaccinia virus* à un niveau acceptable de toxicité.

Comme nos partenaires ne possédaient pas les ressources adaptés pour tester les nouveaux composés sur d'autres virus, nous avons utilisé une méthodologie originale de l'analyse de l'espace chimique, développée dans notre équipe afin de montrer que les composés sélectionnés par criblage pouvaient potentiellement avoir un large spectre d'activité antivirale. En particulier, la carte, c.-à-d. la représentation en deux dimensions, de l'espace chimique occupé par un grand nombre de composés antiviraux permet d'identifier des régions de l'espace chimique peuplées par des molécules actives contre des types particuliers de virus desquels il a été possible d'extraire des motifs structuraux caractéristiques (*privilégiés*). Les deux molécules identifiées expérimentalement à l'ICBFM ont été positionnées sur cette carte dans une région peuplée par des composés anti-MRV (virus à ARN double brin) de la famille des triazolotriazinoindoles. Cette observation permet de présumer que les molécules conçues durant ce travail pourraient posséder une activité biologique similaire, ce qui nécessite toutefois une confirmation expérimentale.

Ce manuscrit est composé de cinq parties. La première présente un compte-rendu de la littérature concernant les cibles de thérapies antivirales, des composés connus efficaces et de précédents rapports d'études de modélisation sur ce sujet. La seconde section décrit les méthodes numériques utilisées dans cette étude. La section 3 est une présentation d'un modèle QSPR de la solubilité aqueuse, faisant parti d'un flux opérationnel pour l'estimation de la biodisponibilité. La section 4 décrit les modèles développés et la procédure de criblage virtuelle qui a conduit à suggérer de nouveaux antiviraux. Finalement, la section 5 est dédiée au développement de la base de données de composés antiviraux, ainsi qu'à la visualisation et à l'analyse de l'espace chimique antiviral.

SECTION 1 COMPTE-RENDU BIBLIOGRAPHIQUE

La première partie fait le bilan de la situation dans le domaine des médicaments antiviraux. Une attention particulière est portée aux composés à large spectre.

La seconde passe en revue les méthodes chémoinformatiques utilisées pour la conception de médicament assistée par ordinateur, tel que le QSAR et la modélisation par pharmacophore, le docking et la recherche par similarité.

SECTION 2 APPROCHES ET OUTILS NUMERIQUES

Cette section décrit les approches chémoinformatiques et les outils utilisés pour ce travail : les relations structures-activités quantitatives (QSAR), les méthodes d'apprentissage automatique (Forêts Aléatoires, Cartes Topographiques Génératives), les descripteurs moléculaires (ISIDA, SiRMS), les pharmacophores (LigandScout), les outils d'analyse de données (KNIME), la recherche de châssis moléculaires (Scaffold Hunter). Certaines informations au sujet de bases de données de petites molécules (ChEMBL, BioinfoDB, PCI) sont aussi mentionnées.

SECTION 3 MODÈLES QSPR POUR ESTIMER LA SOLUBILITÉ ACQUEUSE

Contrairement à de nombreux modèles de la solubilité aqueuse S_w (mol/l) mesurée à température ambiante, notre modèle permet d'estimer le $\log S_w$ dans la gamme de températures 4 – 97°C. Le modèle a été construit sur un ensemble de 421 composés organiques extraits du *Yalkovsky Handbook of Aqueous Solubility Data*. Un modèle Quantitatif de Relation Structure-Propriété (QSPR) a été construit avec l'algorithme des Forêts Aléatoires les descripteurs moléculaires SiRMS. Le modèle est robuste en validation croisée et ses performances sont mesurées par un coefficient de détermination R^2 et une erreur quadratique moyenne RMSE de 0.96 et 0.21 $\log S_w$ respectivement tout

en conservant des performances raisonnables sur un jeu de données externes (RMSE=0.67 log S_w).

SECTION 4 CONCEPTION ASSISTÉE PAR ORDINATEUR D'ANTIVIRAUX À LARGE SPECTRE

Cette section décrit la conception assistée par ordinateur de nouveaux intercalants d'acides nucléiques possédant une activité antivirale incluant les différentes étapes ci-dessous.

Préparation des données. Le jeu de données utilisé pour la construction des modèles contient 167 composés synthétisés et testés au PCI (jeu de données PCI). Chaque molécule contient un fragment plan polycyclique lié à une fonction amine. 117 des 167 composés ayant un effet antiviral maximal E_{\max} (%) connu ont été utilisés pour la construction de modèles QSAR. Les composés ont été répartis dans deux catégories, l'une « active » incluant les antiviraux ayant un $E_{\max} \geq 50\%$ et l'autre « inactive » incluant les composés ayant un $E_{\max} < 50\%$. 161 des 167 composés ayant un fort potentiel intercalant ont été choisis pour développer un modèle pharmacophorique.

Construction et validation des modèles. Trois différents types de modèles ont été préparés : (i) des filtres, (ii) des modèles pharmacophoriques et (iii) des modèles QSAR. Les filtres ont été conçus en utilisant le jeu de données PCI entier. Ils représentent un ensemble de règles définissant des valeurs minimales et maximales pour certains paramètres structuraux : le nombre de cycles fusionnés, le nombre de donneurs et d'accepteurs de liaisons hydrogènes, le nombre de liaisons rotatoires et le poids moléculaire. Seuls les composés satisfaisant à chacun des critères du filtre sont choisis pour le criblage virtuel, les autres étant écartés.

Les modèles pharmacophoriques tridimensionnels ont été développés à l'aide du logiciel LigandScout. Au total, 5 modèles ont été construits. Ils ont été validés sur un jeu de données incluant à la fois des composés actifs et inactifs provenant du jeu de données PCI et d'un échantillon de 20000 composés provenant de la base de données ZINC, considérés comme des leurres (des molécules inactives). Trois modèles ayant une précision > 0.65 ont été choisis pour faire parti du flux opérationnel de criblage virtuel.

Des modèles de classification en deux classes utilisant l'algorithme des Forêts Aléatoires et les descripteurs moléculaires SiRMS ont été construits. Les modèles produits ont des capacités de généralisation raisonnables, permettant d'obtenir un score de précision balancée de 0.74 sur le jeu de données externe.

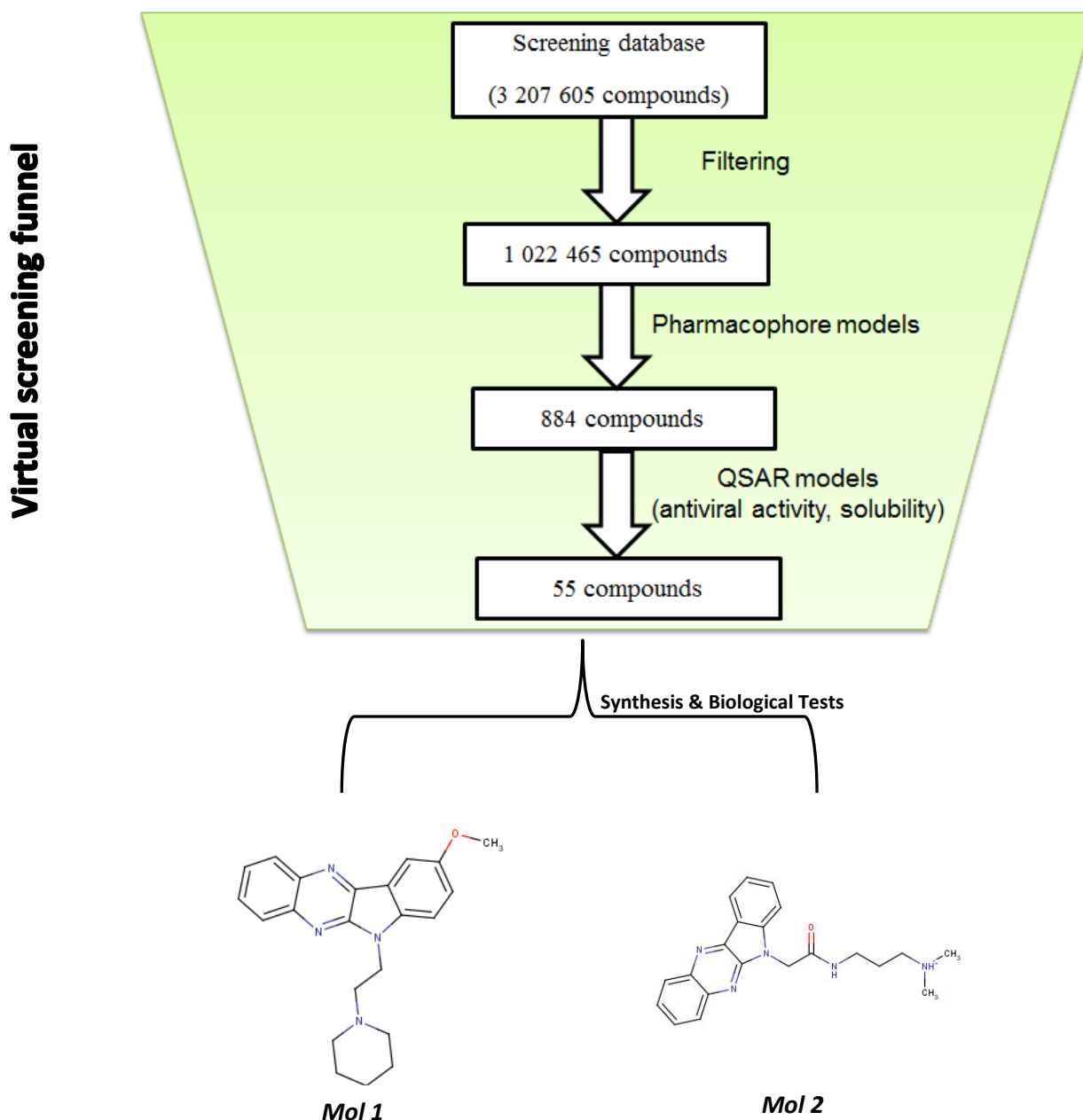


Figure A. Processus opérationnel utilisé pour la conception assistée par ordinateur d'antiviraux intercalants d'acides nucléiques.

Tous ces modèles, ainsi que le modèle de solubilité aqueuse décrit dans la section 3 sont intégrés dans le processus opérationnel de criblage virtuel.

Criblage virtuel. La base de données BioinfoDB, contenant environ 3 millions de structures chimiques disponibles commercialement, auxquelles s'ajoutent 288 composés virtuels générés par combinaison de fragments structuraux typiques d'intercalants provenant de la base de données PCI ont été utilisés pour un criblage virtuel résultant en la sélection de 87 composés touches. Le logiciel PASS n'a pas permis d'identifier de quelconques effets secondaires, ni une toxicité ou une mutagénicité particulière, parmi les composés touches et la base de données PubChem a révélé qu'aucun composé parmi les touches sélectionnées n'a été utilisé précédemment dans une campagne de criblage expérimentale biologique antivirale. 55 composés de cette liste de composés touches ont été synthétisés à PCI puis testés expérimentalement à l'ICBFM. Pour deux composés (*Mol1* et *Mol2*, voir Figure A) les tests biologiques ont montrés une activité significative contre *Vaccinia virus* avec un relativement faible niveau de toxicité. Plus précisément, ces composés (i) réduisent la formation de plaques virales d'un facteur 6 à 8 et (ii) montrent une affinité raisonnable pour l'ADN : les constantes d'affinité mesurées expérimentalement ($\lg K_a$) sont de 6,03 et 5,20 pour *Mol1* et *Mol2*, respectivement. Il faut noter qu'aucun des composés suggérés n'a d'activité contre les virus à ARN simple brin (*Encephalomyocarditis virus*) ni n'induit la production d'interférons cellulaires à un niveau substantiel.

SECTION 5 VISUALISATION ET ANALYSE DE L'ESPACE CHIMIQUE DES COMPOSÉS ANTIVIRAUX

Développement d'une base de données d'antiviraux. La base de données de bio-activités publiquement accessible ChEMBL a été utilisée comme source de données sur des composés antiviraux. La curation de données a été entreprise en utilisant le logiciel KNIME et la standardisation des structures chimiques a été réalisée à l'aide du logiciel Standardizer édité par la société ChemAxon. Les données d'activité ont été conservées si elles étaient publiées dans un journal scientifique et n'étaient pas annotées comme non-valide ou comme doublon. Seules des données quantitatives dans un intervalle de valeur défini ont été conservées. Au total, 24629 composés ont été sélectionnés pour la base de données antivirale. La classe d'activité antivirale a été définie par le type de pathogène viral contre lequel un composé a une activité. Ceci a conduit à choisir les

catégories suivantes : Enterovirus (424), Hepacivirus (5320), Influenza A (638), Lentivirus (8854), Orthohepadnavirus (700), Pestivirus (412), Simplexvirus (790) and autres antiviraux (7897).

Modélisation GTM. Une carte topographique générative (GTM) est une méthode de réduction de dimensionnalité représentant des objets à partir d'un espace de descripteurs moléculaires initial de grande dimensionnalité vers un espace latent de deux dimensions qui peut être décrit par une grille rectangulaire. Le grand avantage de la GTM est le calcul d'une distribution de probabilité des données qui peut être ensuite exploitée pour une modélisation structure-activité.

Le logiciel ISIDA/GTM et les descripteurs moléculaires fragmentaux ISIDA ont été utilisés pour construire les modèles GTM utilisés pour l'analyse des caractéristiques structurales des composés de la base de données d'antiviraux. L'ensemble des descripteurs moléculaires les plus appropriés a été choisi parmi ceux qui ont conduit aux modèles de classification les plus performants pour distinguer des composés « actifs » des composés « inactifs » pour chacune des sept classes d'antiviraux au cours d'une validation croisée en 3 paquets (ayant un score de précision équilibrée > 0.7). De cette façon, les trois « meilleurs » modèles GTM, chacun décrivant un espace de descripteurs moléculaires particuliers, ont été conservés.

A l'étape suivante, 1,2 millions de composés de la base de données ChEMBL sans aucune annotation sur leur activité antivirale ont été positionnés sur les cartes GTM. Les cartes ont été colorées en utilisant un code couleur selon la catégorie « active » ou « inactive » des composés majoritairement présents sur les nœuds de la grille, pour mieux visualiser les régions les plus saturées en antiviraux.

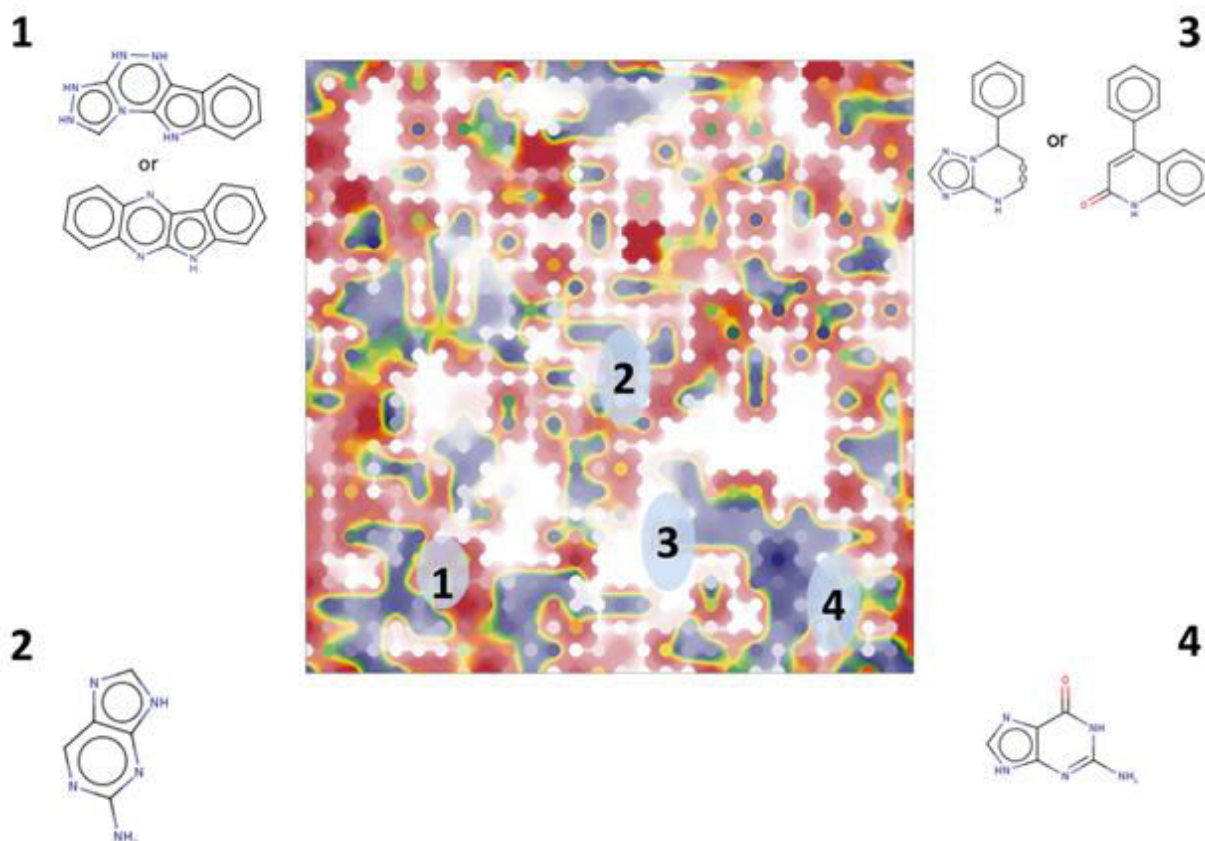


Figure B Carte GTM construite en utilisant les composés antiviraux comme jeu d'entraînement. Le code couleur montre les régions du paysage chimique où sont situés les composés antiviraux (en bleu) contre les composés sans activité antivirale (en rouge). Les nombres sur la carte accentuent les régions où sont localisées des motifs structuraux privilégiés (PSMs).

Identification de caractéristiques structurales. Une des propriétés de la GTM est qu'une molécule localisée en un point de l'espace latent peut aussi être délocalisée sur plusieurs nœuds de la grille couvrant la carte, chacun étant associé à la probabilité (aussi appelée *responsabilité*) de présence de la molécule dans l'espace initial. L'ensemble des valeurs de responsabilité est identifié aux composantes d'un vecteur, appelé *motif de responsabilité* (RP), qui est unique pour une molécule donnée. Les molécules ayant un motif de responsabilité similaires sont localisées dans les mêmes régions sur la carte conformément à leurs positions dans un même voisinage de l'espace des descripteurs initial.

Les caractéristiques structurales des composés antiviraux ont été discutées en terme de châssis moléculaires et de motifs structuraux privilégiés (PSMs). Les châssis

sont définis comme une sous-structure commune partagée par les composés d'une certaine classe antivirale, incluant au moins 3 fragments cycliques – fusionnés ou interconnectés par des ponts acycliques. Un PSM est composé de sous-structures, connectées ou non, partagées par les composés d'une ou plusieurs classes d'une part, et d'autre part, qui apparaissent rarement (ou sont absents) des autres classes ou dans les composés qui n'ont pas d'activité antivirale. Ils sont établis à partir de l'analyse d'ensembles de molécules possédant des RP similaires. Au total, huit PSMs ont été identifiés, tels que ceux illustrés sur la Figure B.

Les motifs de responsabilités de *Mol1* et *Mol2* sur la carte GTM sont similaires aux RPs des indolequinaxolines et des triazolotriazinoindoles présents dans le jeu d'entraînement. Les indolequinaxolines sont connus pour être actifs contre le *Vaccinia virus* et le CMV qui sont des virus à ADN double brin, tandis que les triazolotriazinoindoles sont actifs sur le virus ARN à double brin *Mammalian orthoreovirus*. Ceci laisse donc supposer que *Mol1* et *Mol2* auraient une activité contre ces virus.

CONCLUSIONS

1. Un ensemble d'outils de modélisation incluant des filtres, des pharmacophores et des modèles QSAR qui ont été développés dans ce travail, ainsi que l'estimation de risques d'effets secondaires et de propriétés ADME/Tox basés sur des logiciels commerciaux ont été utilisés pour le criblage virtuel d'une base de données de plus de 3M de composés. Les touches sélectionnées ont été synthétisées et testées expérimentalement par nos partenaires. Les expériences ont montré que deux des molécules proposées virtuellement ont une activité réelle forte sur des virus à ADN double brin pour un niveau de toxicité acceptable.

2. Une base de données exhaustive sur les antiviraux référencant 24629 composés a été assemblée. Les structures chimiques ont été curées et annotées selon leur activité contre certains *Genus* viraux.

3. Les composés de la base d'antiviraux ont été analysés au moyen de cartes génératives topographiques (GTM) et par l'étude d'agrégats de châssis structuraux. Les GTMs révèlent plusieurs régions compactes peuplées par des composés antiviraux appartenant aux mêmes chémotypes. L'analyse des composés de ces zones a permis d'identifier des motifs structuraux propres à certains types d'activité antivirale. Les deux composés identifiés par criblage

virtuel et confirmés expérimentalement ont été localisés dans des zones de la carte peuplées par des agents anti-MRV, ce qui conduit à supposer que ces molécules auraient une activité antiviral sur les virus à ARN double brin.

Un nouveau modèle QSPR estimant la solubilité aqueuse à différentes températures a été développé. Il a été utilisé pour estimer la solubilité des touches sélectionnées à différentes étapes du criblage virtuel.

INTRODUCTION

Viral infections are known to be a cause of many dangerous diseases. Recent outbreaks of Influenza A virus (USA, 2009) [1] and Ebolavirus (Liberia, 2013-2015) [2] were devastating, resulting in many casualties and deaths in the population making the search for new antiviral drugs a crucial task.

Virus reproduction consists of several stages. Thus, many strategies can be used to tackle the problem. Existing antiviral drugs mostly target specific viral proteins which provide inhibition of particular virus reproduction. Inhibiting attachment proteins or reverse transcriptases has the lesser risks of “collateral damage”, i.e. negative impact on host cells due to drugs unspecific binding to cells proteins. Above-mentioned types of proteins are not present in cellular organisms and, therefore, cell life cycle is usually not affected. However, this methodology does prevent the emergence, by mutation/selection, of drug-resistant strains – for example, against non-nucleoside reverse transcriptase inhibitors (e. g. efavirenz [3]). Also, virus-specific drugs are by definition not useful as broad-spectrum antivirals.

Another antiviral drug development strategy is a search for compounds with broad spectrum of antiviral activity. Major groups of broad-spectrum antivirals include nucleotide analogs (e. g. acyclovir) and small molecule interferon inducers (e. g. tilorone [4]). Nucleotide analogs are interfering with virus transcription [5,6], whereas interferon inducers are believed to enhance hosts immune response [6] and disrupt virus protein translation [7]. However, nucleotide analogs mostly display activity against DNA viruses and retroviruses and due to their low oral bioavailability and significant toxicity are of limited value in the treatment of chronic diseases, for which the oral therapies are highly desired. Small-molecule interferon inducers can exhibit activity against both RNA and DNA viruses but they lack effectiveness in complete eradication of viral infections, whereas interferon itself is quite expensive and has strict requirements in terms of storage and distribution.

One of the underestimated classes of broad-spectrum antivirals are nucleic acids intercalators. Intercalation changes the conformation of viral nucleic acids leading to inability to fulfill their biological function, preventing viruses from replication, and, possibly, distorting R (D)NA-dependent RNA-transcriptase interaction with viral nucleic acid. Nucleic acid intercalators have already been tested against certain viruses *in vitro* [8,9] and some of them were found active. The fact that intercalators target nucleic acids

instead of proteins gives a certain expectation that problem of high mutation rate of virus, and, thus, rapidly developing resistance can be sorted out. This makes nucleic acid intercalators particularly interesting for further antiviral drug development.

The goal of this study is computer-aided design of new intercalators possessing broad-spectrum antiviral activities. We used various chemoinformatics approaches (structural filters, QSAR and pharmacophore) to build predictive models used in virtual screening of a databases containing more than 3M compounds. Virtual screening resulted in selection of 55 hit compounds which first were synthesized at the A.V. Bogatsky Physico-Chemical Institute (PCI) NAS of Ukraine in Odessa, Ukraine and then experimentally tested at the Institute of Chemical Biology and Fundamental Medicine (ICBFM) in Novosibirsk, Russia. Two indolequinaxoline derivatives were found to be DNA intercalators and active against *Vaccinia virus*, at an acceptable level of toxicity.

We used original methodology of chemical space analysis developed in University of Strasbourg Chemoinformatics laboratory in order to show that compounds selected in screening may have broad-spectrum antiviral potential. Firstly, a large dataset of antiviral compounds extracted from the ChEMBL database was curated and annotated according to compounds activity against a certain virus *Genus*. Secondly, Generative Topographic Mapping (GTM) was used for chemical space analysis, providing identification of the zones in chemical space populated by actives against particular type of viruses from which characteristic “privileged” structural patterns could be extracted. Moreover, the GTM model allowed predicting activity against major virus *Genus* for projected compounds. Unfortunately, experimental validation of the assumed broad-spectrum antiviral activity, pending collaboration with dedicated antiviral screening facilities, could not be achieved within the timeframe of the PhD thesis.

The manuscript consists of five parts. The first one represents a literature review on targets of antiviral therapy, known effective compounds and previously reported modeling studies. The second section describes computational methods used in this study. Part 3 reports QSPR model for aqueous solubility developed as part of this work virtual screening workflow for bioavailability assessment. Part 4 describes models development and virtual screening procedure which resulted in suggestion of new antivirals. Finally, Part 5 is dedicated to development of antiviral database, as well as visualization and analysis of antiviral chemical space

PART 1 REVIEW ON VIRUS PROBLEMATIC

1.1 Overview of the virus structure and reproduction

A virus is an invasive biological agent, one of the smallest among the enormous variety of life forms. -Viruses are bound to reproduce inside the cells of living hosts and, therefore, depend on cell structure and metabolism.

A virus has several determining features. For example, a virus has only one type of nucleic acid - either DNA or RNA, whereas other living organisms have both. Another unique feature of viruses is absence of the protein synthesizing system. It has to use its host's system in order to reproduce, particularly by introducing its genetic information to the cell. This is a very specific form of parasitism – the genetic parasitism. Viral reproduction is thus a self-assembly process of disjoined components produced by the host viral genome-infected cell machinery [10].

Outside the living cell, a virus exists in the form of a virion, which consists of genetic material and compounds that keep genetic material unharmed [11]. Virions consist of two types of compounds: primary and auxiliary. Primary compounds are present in all types of viruses and they are crucial for virus existence. Nucleic acids and proteins are primary compounds. The variety of nucleic acids forms in viruses is incredible: unlike cell organisms, virus genome can be represented by both DNA and RNA, which could be either single-stranded or double-stranded with linear or circular molecular shape. Proteins can be further classified in structural and non-structural. Structural proteins form capsid – a special type of protein coat which protects nucleic acids from decomposition due to interaction with nucleases. Also, some viral proteins are covalently bonded with nucleic acids and play the role of terminal proteins; in this case protein-nucleic acid structure is called nucleocapsid. Non-structural proteins are either enzymes, which play a role in the virus reproduction or regulatory proteins, which define the beginning of reproduction process, the end of reproduction process, etc. Most common auxiliary compounds are lipids or carbohydrates in glycoproteins. Lipids are the main constituent of envelope, which is a specific viral formation similar to cells membrane. Glycoproteins are located on the envelope or outer part of capsid and their role in virus reproduction is to capture specific receptors located on cells surface.

Viruses display features of the living organism only inside the host cell. The interaction between virus and cell consists of 7 stages (Figure 1): 1. Attachment 2.

Penetration 3. Uncoating 4. Transcription 5. Translation 6. Genome replication 7. Self-assembly and 8. Release

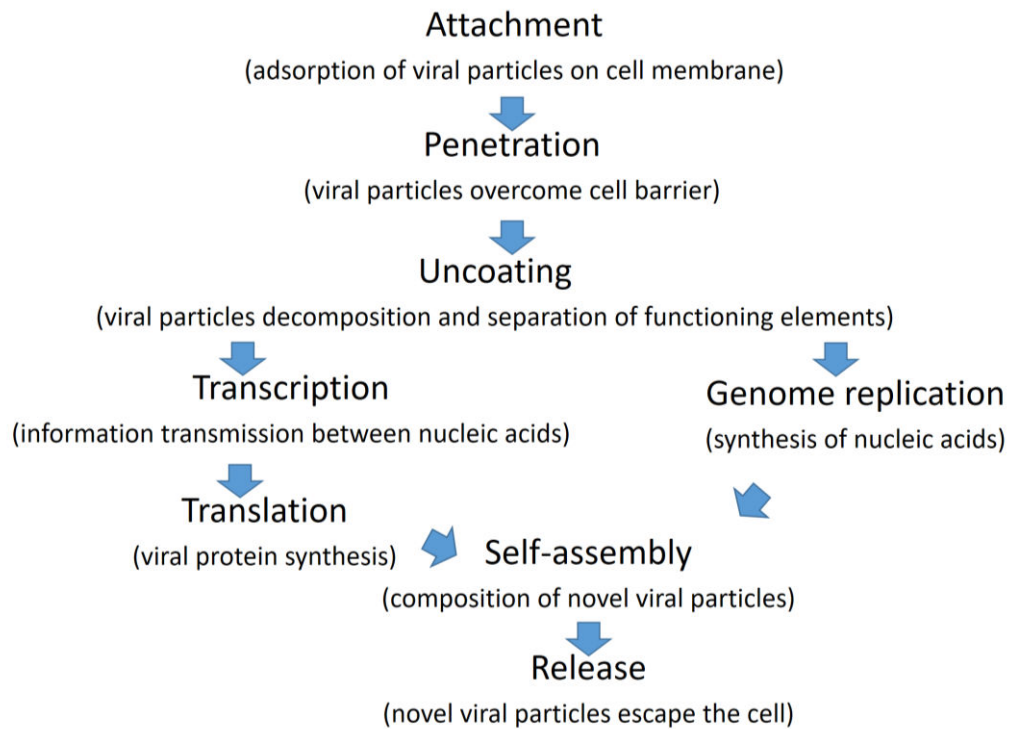


Figure 1. Stages of viral reproduction

In order to infect the cell, the virion must bind to the cell surface and uncoat itself so that its genome becomes accessible for the cell system for viral transcription or translation [11]. This adsorption process is called ‘attachment’: a specific binding between viral attachment protein and specific receptors on the host cellular surface (example [12]). This process might begin as an unspecific electrostatic attraction between above-mentioned parts, however further interaction requires specific binding between cell surface receptors and viral attachment proteins. Viruses use cell surface receptors for penetration. Some, like *Vaccinia virus* can even have multiple types of surface proteins [13]. Binding to only one receptor species is not enough for cell penetration, in order to entry virus must bind to a sufficient number of receptors which leads to inevitable changes of the cell membrane structure [11]:

Cell penetration occurs almost immediately after attachment. Penetration follows one out of three possible scenarios: 1) by membrane fusion [14] 2) by endocytosis [10] and 3) pore-mediated [15]. Most viruses, enveloped or not, enter the cell via endocytosis [16]. Membrane fusion is only feasible by enveloped viruses, and pore-mediated

penetration is intrinsic to non-enveloped viruses. Endocytosis provides intracellular transport for the viral particle as a part of endocytic vacuole, because it can move in any direction and fuse with any cell membrane including the one of the nucleus . This would allow virus particles to infect any cell organelle. Envelope-membrane interaction results in total fusion allowing viral genome to end up inside the cell (Figure 2). Non-enveloped viruses can interact with the membrane by means of their capsid protein, which leads to formation of pores used by virus in order to get inside.

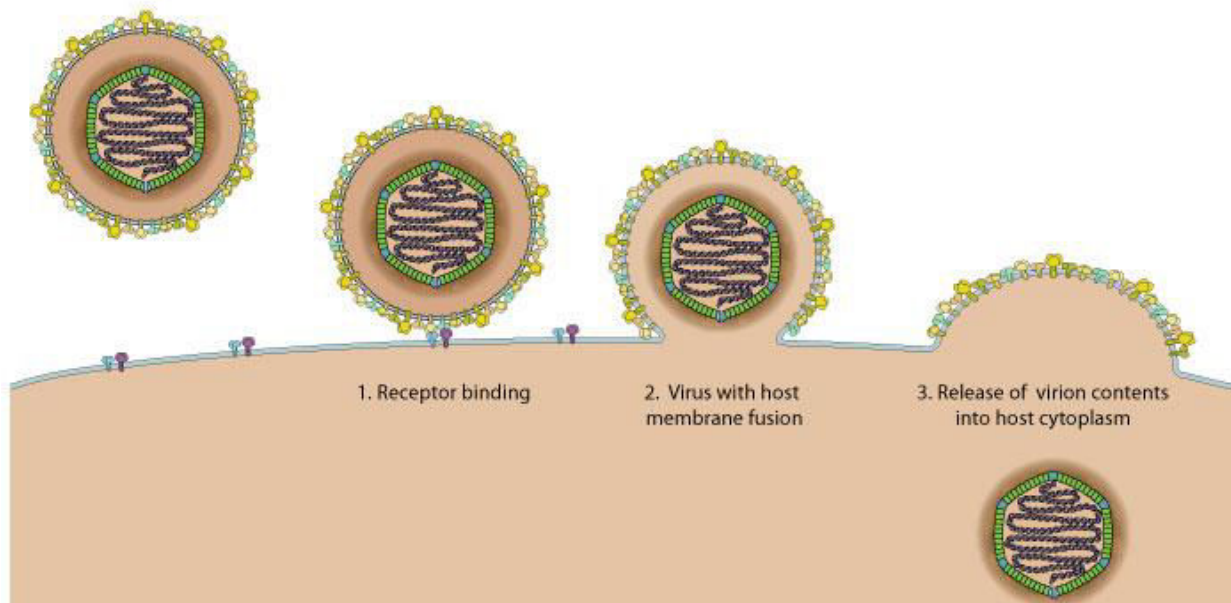


Figure 2. Schematic representation of attachment, penetration and uncoating stages [17]

Uncoating is crucial for the virus because it allows the viral genome to express itself. In this process, the viral capsid is removed, possibly due to degradation caused by viral enzymes or host enzymes or due to simple dissociation. This releases viral genomic nucleic acid. Sometimes the whole infection process is defined by whether virus is able to uncoat inside this particular cell. For viruses with helical symmetry, such as Influenza A, the uncoating does not lead to the total removal of all capsid proteins, because some of them are needed in order to form nucleocapsid [18]. Notice that uncoating and intracellular transport are related processes. If intracellular transport does not work properly at the site of uncoating, the viral particle ends up in the lysosome and, therefore, can be decomposed by lysosomal enzymes.

According to a central dogma of molecular biology [19] the order of nucleotides determines the protein structure. For DNA viruses, protein synthesis follows the classical path (Figure 3) involving DNA to RNA transcription [20]. In case of RNA viruses, there are

several options. Some viruses of this kind (Positive-sense viruses) may use their nucleic acid for direct translation [21]. Their genome is able to infect the cell and to be used for translation at once, without creating the messenger RNA. Another type of sequential information transfer requires both nucleic acid and protein in form of nucleocapsid to be present inside the host cells. In this case, initial RNA is used as matrix for complementary RNA synthesis and synthesized RNA is used in translation [22]. The third option for information transfer by RNA virus is to use initial viral RNA to create DNA via reverse transcription process. After some permutations, viral DNA integrates with host cell DNA then creating viral RNA used for protein synthesis.

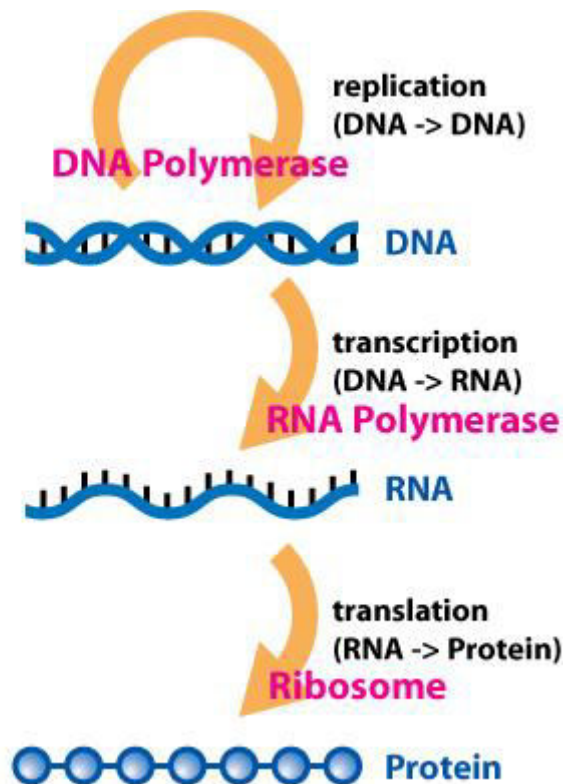


Figure 3. Schematical representation of the sequential information transfer [23]

The other key aspect of viral reproduction is replication, *i.e.* multiplication of the genome by synthesis of its nucleic acids. Viral genome replication is mediated by regulatory protein expression. In a double-stranded DNA virus the replication mechanism is the same as for eukaryotes, whereas single-stranded RNA viruses simply use RNA polymerase to copy the initial viral RNA [10].

The basis of the self-assembly process is specific protein-nucleic acid recognition, which might occur as a result of hydrophobic, ionic, hydrogen binding or spatial match. Protein-nucleic acid recognition happens at specific nucleotide sequence in the coding

part of the genome. This part of nucleic acid is a starting point of viral particle assembly, which continues due to specific protein-protein interaction. Self-assembly requires a virus to form a so-called virus machinery and to use substances from the host cell to complete the final assembly of a new viral particle [24].

A virus can leave the cell according to two different mechanisms. The first one (lysis) is a process that kills the cell by bursting its membrane and cell wall if present. The second one (budding) allows the virus to be wrapped up by the cell membrane in order to be released. Prior to budding, the virus may place its own receptor onto the surface of the cell, in preparation for the virus to bud through, forming an envelope with the viral receptors already on it. The second type of release is more common among viruses because it allows them to keep the cell alive and reproduce until the cell is totally exhausted [11].

Just like any other organisms, viruses are prone to mutations. Their relative simplicity actually allows for very high mutation rates. Therefore, viral proteins' structure is changing fast – which makes them an uncomfortable target for antiviral therapy. Mutation mechanisms can also vary (Figure 4), i.e. deletion, insertion or substitution of certain nucleotides and the process itself can be also divided in two groups: spontaneous and induced [25,26].

Induced mutations are caused by the direct impact of various mutagens of different nature. These mutagens usually belong to two classes: chemicals and radiation. Among chemical mutagens one can distinguish base analogs, intercalators, alkylating agents, etc. [27,28]. Induced mutations allow virus to change its proteins structure, thus avoiding inhibiting properties of antiviral compounds. Even though, the probability of nucleotide substitution per strand copying is rather small (8.9×10^{-6}) [29], the high frequency of virus reproduction [30] may lead to quick formation of numerous drug resistant offspring.

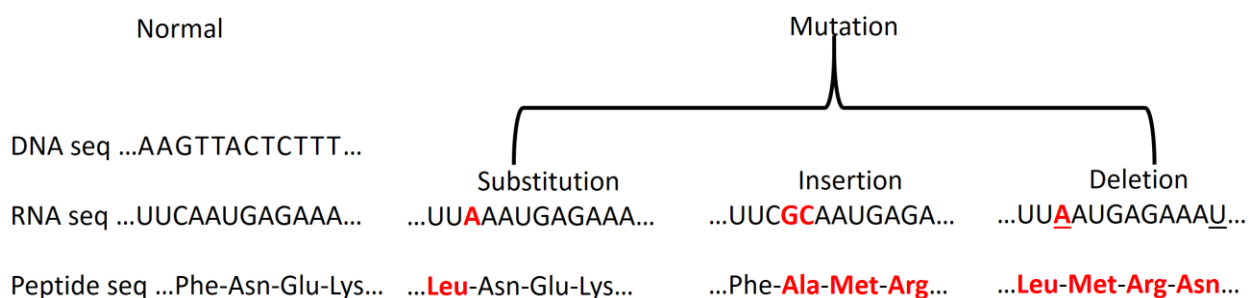


Figure 4. Schematic representation of the mutation types

Viral pathologies are a straightforward consequence of virus impact on infected cells. These changes are: *i)* physical damage of the cell components and alteration of physico-chemical parameters (pH, membrane viscosity); *ii)* lysosome dysfunction which results in uncontrolled lysosome enzymes liberation leading to cell autolysis; *iii)* intensive depletion of cell protein synthesis resources due to virus reproduction; *iv)* destruction of specific molecules in the cell [11].

The impact on the-level of the whole organism varies according to certain criteria: *i)* infectious process duration; *ii)* symptoms; *iii)* spreading of the infection. One of the most prominent examples of a devastating viral disease is smallpox [31]. In the 1967, The World Health Organization (WHO) estimated that 15 million people contracted the disease [31].-The disease is highly contagious and has an airborne transmission path. Nowadays smallpox is believed to be eradicated due to stepwise preventive measures.

The representative example of modern highly dangerous viral disease is AIDS. It is caused by HIV and leads to the significant decrease in immune system activity, especially non-specific immune response. Therefore, people with AIDS are prone to die from other less harmful infections [32]. Another example are oncoviruses, such as Kaposi sarcoma virus, which cause tumor formation. WHO International Agency for Research on Cancer estimated that in 2002 17.8% of human cancers were caused by viral infection [33].

Although above-mentioned diseases are deadly, it is worth mentioning that viruses are also responsible for causing many less dangerous diseases. For instance, many respiratory diseases have viral origins. Viral pneumonia occurs in about 200 million people a year, including approximately 100 million children [34].

1.2 Current antiviral treatment strategies

Strategies of viral disease treatment remedies can be divided in two major groups depending on their nature: vaccination and drug therapy. Vaccination is the administration of antigenic material (a vaccine) to stimulate an individual's immune system to develop adaptive immunity to a pathogen. This antigenic material may consists of inactivated or attenuated viruses, or use artificial epitopes or virus-like particles [35,36,37]. Inactivated vaccines are used against poliomyelitis [38], rabies [39] and influenza [40]. The success of these vaccines is controversial. While polio can be successfully prevented, influenza vaccine gives only partial protection [41]. Attenuated vaccines require live virus or bacteria strains with very low virulence. These vaccines may be produced by passaging, for example, adapting a virus into different host cell cultures, such as mammalian cells, or at suboptimal temperatures, allowing selection of less virulent strains, or by

mutagenesis or targeted deletions in genes required for virulence. Vaccines made of virus-like particles consist of protein derived from the structural proteins of a virus. Therefore, human organism will be familiar with viral antigens in case the real outbreak occurs. No matter what the strategy of vaccine development is, there are some viral diseases which cannot be prevented by vaccination [42], at least for now.

Taking this into account, as well as the fact that vaccination is mostly a preventive tool against viruses, the need of antiviral drugs is obvious. The applicability domain of existing antiviral drugs is limited. The WHO report on essential pharmaceuticals [43] gives a list of most important antiviral drugs which should be provided as a part of basic health care. The list is divided in terms of activity into following categories:

- Antiherpes medicines (acyclovir)
- Antiretrovirals (abacavir, nevirapine, indinavir etc.)
- Other antivirals (oseltamivir, ribavirin, interferon alpha)

This classification, however, does not reflect antiviral mechanisms of action. Worldwide-known acyclovir alongside with several antiretrovirals (Figure 5) belong to a group of so-called nucleoside analogs [6]. These compounds resemble DNA or RNA nucleosides and can potentially be captured by enzymes or tRNA involved in virus reproduction. This might lead to synthesis of a non-coding sequence in viral nucleic acids. Even though this activity mechanism is not specific to one virus, nucleoside analogs are effective only against particular virus strains [5].

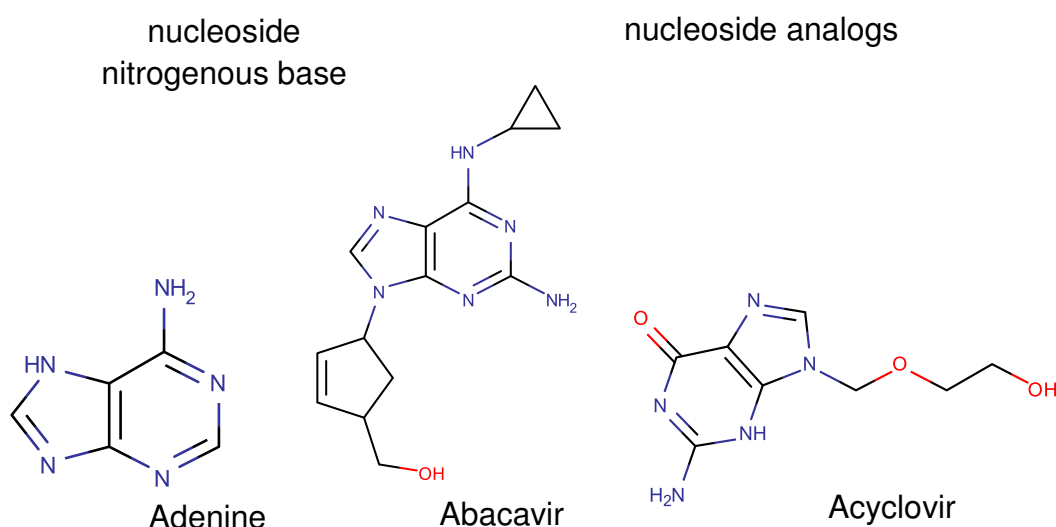


Figure 5. Example of nucleoside analogs

Another essential category is viral proteins inhibitors (Figure 6). Targeting proteins, unlike targeting DNA/RNA, is the common drug design strategy, and can be supported by protein structure determination/ structure-based drug design [44,45]. Since viral reproduction can be broken at different stages, technically, any type of viral protein can be used as antiviral drug target.

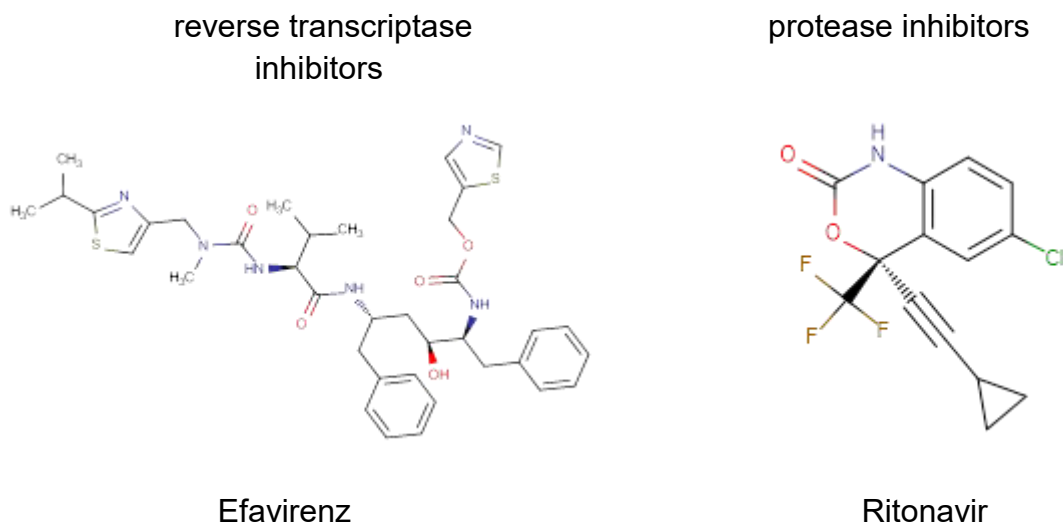


Figure 6. Example of HIV viral proteins inhibitors

Viral protein inhibitors – as well as vaccines – are prone to be rendered obsolete by the emerging of resistant mutants. These mutations may actually be enhanced by the destructive effect of the antiviral drug, as a part of virus defense mechanism [46]. Therefore, the structure of viral proteins can change significantly in a short period of time, making new generations of viruses resistant to treatment by protein inhibitors.

The last notable category is the one based on the interferon-based defense mechanism (interferon and interferon inducers). Interferons are a group of proteins [6,47] produced as immune response. It induces synthesis of protein kinase which phosphorylates initiation factor of translation and, therefore, prevents viral proteins from being created [48]. Pure interferon is used in the form of IFN- α . IFN- α is used for the immune boost in various diseases, including HBV [6]. However, it is an expensive drug with short keeping time. As for interferon inducers, they were not included into the WHO report on essential pharmaceuticals but compounds like Tilorone (Figure 7) have been officially approved for use in some countries [49].

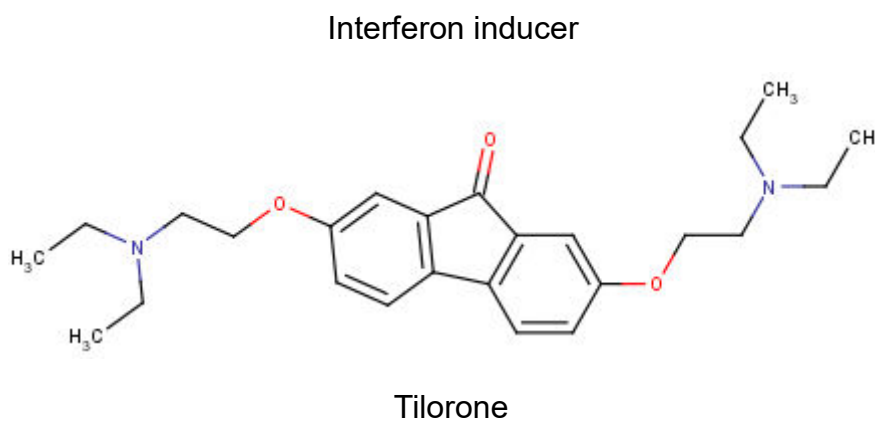


Figure 7. Example of drugs enhancing interferon activity

This compound has been used as broad-spectrum antiviral drugs, although it does not display antiviral activity against all types of viral pathogens [50].

1.3 Earlier studies on computer-aided design of antiviral drugs

There were attempts to use chemoinformatics for antiviral drug design. Langer et al. [51] carried out virtual screening aimed to select HRV coat protein inhibitors. It involved 3 stages: pharmacophore model, docking and similarity search. They used 30 pdb entries of the HRV coat protein in complex with inhibitors from Brookhaven Protein Databank. The first step included pharmacophore model development using the Catalyst program. All inhibitors possess a rather hydrophobic character matching the lipophilic environment of this binding site, and most of them tend to form hydrogen bond to the amide nitrogen of Leu 100 of the viral proteins. Docking calculations have been performed using the LigandFit tool implemented in the Cerius2 software. The ligands were treated as being flexible during docking while the protein was kept rigid. Principal Component Analysis (PCA)-based clustering was applied to assess the hits similarity.

In order to select potent antivirals Maybridge DB containing approximately 60 000 chemical compounds was used for virtual screening. As a result, 6 compounds were selected for in vitro antiviral (anti-HRV) study. The HRV 3C protease inhibitor rupintrivir served as a positive control. Maybridge substance 20 (Figure 8) cells exhibited activity at 10 mg/L while having CC_{50} of 32 mg/L. Furthermore, compound 15 was active at 100 mg/L while having $CC_{50} > 100$ mg/L. Since these two structures display the most beneficial ratios between inhibitory activity and cellular toxicity, they were considered the most

promising antivirals. The positive control displayed inhibitory activity at 1 mg/L with CC_{50} higher than 100 mg/L.

Gao et al. [52] tried to use QSAR methodology and docking for anti-influenza A drug design. The X-ray crystal structure of Influenza virus neuraminidase complex with zanamivir and antiviral activity data on 35 flavonoid compounds were used. Ligand-based pharmacophore models, atom-based QSAR models using partial least squares (PLS) were developed and molecular docking into neuraminidase was made in order to define flavonoids structural features contributing to their antiviral activity. Substituents of aromatic rings with positive and negative contribution to activity, as well as key physicochemical features, were defined. There are more examples of studies dedicated to the virtual screening of antiviral compounds [53,54] with different outcomes, however they all have one thing in common: they were aiming for the design of specific protein inhibitors.

These studies show that significant progress in antiviral compounds development can be achieved using methods of chemoinformatics. However, there are still many challenges and opportunities for improvement. For example, in [52] researchers did not test their hypothesis on improvement of compounds activity by modifying certain structural elements, synthesizing and testing new compounds which would have validated their findings. In [51] new compounds were retrieved and tested with two of them displaying substantial antiviral activity - enough for lead compounds to be further optimized, but not enough for a drug candidate.

A potential drug candidate must not only possess high activity but also have acceptable ADME properties. One of the most important ADME properties is aqueous solubility, since insoluble compounds cannot be even tested in cell-based antiviral bioassays. This property is also one of the main criteria in Biopharmaceutics Classification System which is used to differentiate drugs [55,56]. There were several attempts to apply methods of chemoinformatics to predict aqueous solubility. Quantum chemistry-based approaches [57] showed acceptable results but turned out to be time-consuming, with the rise of prediction error as molecular complexity increases. Therefore, empirical descriptor-based QSPR models became a more popular choice as predictive tools for solubility. Several studies [58,59,60] were carried out using several data source (AQUASOL, PHYSPROP database), different machine-learning techniques (MLR, ANN, PLS) and various molecular descriptors (RDF code values [58], functional groups counts

[59], E-state indices [60]). Models given in these studies showed acceptable prediction capacity on 21 important organic compounds which were part of “solubility challenge” [61]. All QSPR solubility models face two problems: training data accuracy and training set chemical space coverage (compound diversity) [62]. There are several protocols for quantitative aqueous solubility determination, which can lead to different results. Furthermore, solubility may be highly temperature-dependent (for instance, solubility of adipic acid, which is used as excipient for pharmaceuticals [63], increases more than two times in the range from 20 to 40 °C [57]) but this is very often ignored in training data compilations. As for diversity/coverage, data used for model development may have a particular focus (e. g. drug-oriented, popular chemicals-oriented) depending on database it comes from.

Even though these types of problems can occur in any solubility model, there is another rarely addressed issue. Currently available models predict solubility in a quite narrow temperature range (typically 20– 30 °C) [58,59,60], disregard the fact that solubility is a temperature dependent property.

The only work which was dedicated to prediction of solubility at different temperatures using QSPR models so far was described in [64]. In that study temperature was used as a descriptor in Wavelength Neural Network model for prediction of a solubility of 25 anthraquinone dyes in supercritical carbon dioxide at 18-150 °C. However, QSPR model capable of aqueous solubility prediction of structurally diverse organic compounds has not been developed yet.

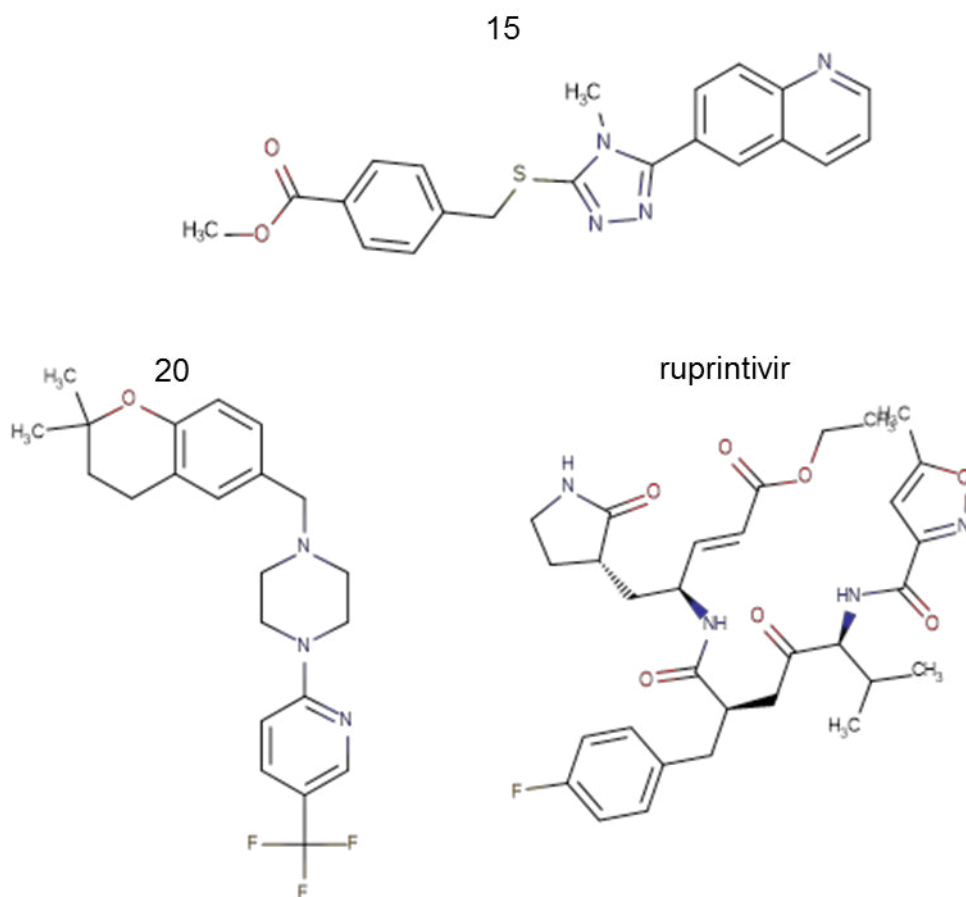


Figure 8. Virtual screening ‘hits’ and control compound

So far, to our knowledge, chemoinformatics and modeling were typically applied for viral protein inhibitor design and, implicitly, for the *in silico* profiling of ADME properties of antiviral drug candidates. This work will, on the contrary, follow a more general approach to antiviral compound design, following an audit of existing antiviral structure-activity information in public databases, and the herewith resulting cartography of relevant “antiviral” chemical space. The contribution specifically features:

- 1) A first attempt for chemical space description of antiviral compounds and computational assessment of suggested virtual screening hits promiscuity.
- 2) An approach for virtual screening of broad-spectrum antivirals, contrary to highly specific single target effecting compounds.
- 3) A QSPR model which predicts solubility within a wide range of temperatures.

PART 2 COMPUTATIONAL TECHNIQUES USED IN THIS STUDY

Computational techniques were used for database development, chemical space analysis and virtual screening tools creation. Virtual screening is an *in silico* analogue of biological screening. The aim of virtual screening is to select compounds with the optimal structures among potential drug candidates, using one or more computational procedures [65]. Virtual screening can be used to choose both compounds from chemical libraries and structures of yet non-existing substances to be synthesized. The tools for the virtual screening can vary, in this work we used QSA(P)R and pharmacophore models.

2.1 (Q)SA(P)R approach - The (Quantitative) Structure-Activity (Property) Relationship

The (Quantitative) Structure-Activity (Property) Relationship approach can be described as an application of data analysis and statistics for development of the models capable of effective quantitative prediction of compound properties or biological activities based on their structures. Model development is based on three key elements: (1) a dataset providing both compounds chemical structures and experimental values of their biological activity or property; (2) molecular descriptors needed for mathematical representation of structures; and (3) machine-learning algorithms for determination of relationship between structures and activity [66].

Fragment-based molecular descriptors were mainly used in this study.

Two different fragmental approaches for representation of molecular structure at 2D level have been used: Simplex representation (SiRMS) [67,68] and ISIDA descriptors – Substructure molecular fragments (SMF) [69]. These descriptors are proved to be effective for QSAR task solving [70,71,72]

2.1.1 SiRMS (Simplex representation of molecular structure)

Two-dimensional (2D) simplexes [67,68,70] are four-atom fragments with fixed composition and topology. Simplexes are called “bounded” if all vertices are connected.

The descriptor vector is defined as the number of occurrences of each simplexes in a molecule. Simplex vertices are labeled according to various characteristics of corresponding atoms. Apart from elements, different physico-chemical characteristics of atoms can be used for atom labeling in simplexes, e.g. atom types, partial charge, lipophilicity, refraction, interatomic potentials and donor/acceptor propensity in hydrogen-bond formation. For continuous atom properties the change of numerical data into ordinal

values (Figure 9): (i) partial charge $A \leq -0.05 < B \leq 0 < C \leq 0.05 < D$, (ii) lipophilicity $A \leq -0.5 < B \leq 0 < C \leq 0.5 < D$, (iii) polarizability $A \leq 1.5 < B \leq 3 < C \leq 8 < D$, (iv) VDW attraction $A \leq 50 < B \leq 100 < C \leq 250 < D \leq 400 < E \leq 650 < F \leq 2000 < G$, (v) VDW repulsion $A \leq 20000 < B \leq 32000 < C \leq 50000 < D \leq 100000 < E$, (vi) Lennard-Jones distance $A \leq 0.05 < B \leq 0.1 < C \leq 0.2 < D \leq 0.3 < E \leq 0.5 < F$, (vii) Lennard-Jones energy $A \leq 2.5 < B \leq 3 < C \leq 3.5 < D \leq 4 < E$ and (viii) electronegativity $A \leq 2.19 < B \leq 2.5 < C \leq 3 < D$. H-bond formation potential is indicated as A (acceptor of hydrogen in H-bond), D (donor of hydrogen in H-bond), and I (indifferent atom).

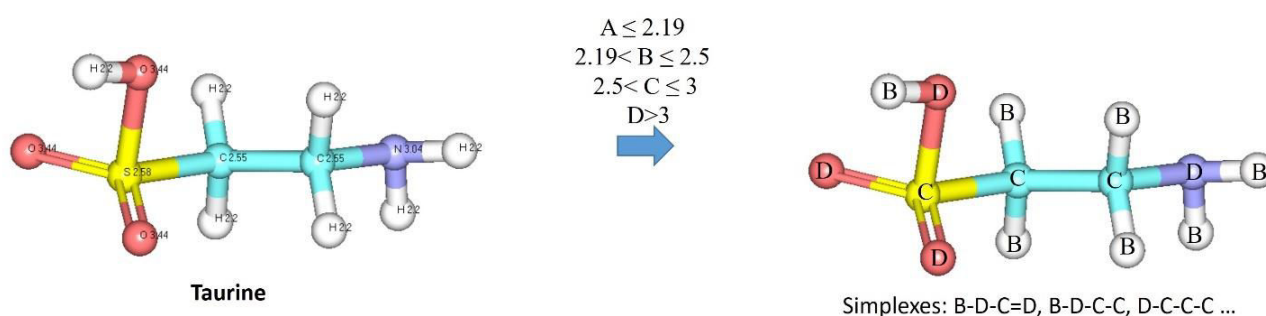


Figure 9. Example of electronegativity-labeled simplex generation

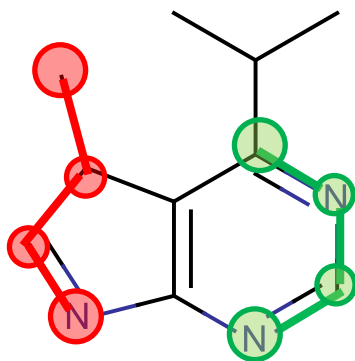
Even though atom labeling and descriptor generation algorithm for SiRMS descriptors was originally used on simplexes, up-to-date HiT QSAR software [67] can carry out this procedure for fragments with sequence size from 1 to 10. This software is also capable to calculate other descriptors which take into account integral characteristics of the molecule, such as molecular weight, lipophilicity etc.

2.1.2 ISIDA substructure molecular fragments

The Substructure molecular fragment algorithm [69,73] was used in ISIDA/QSPR software for molecular fragments generation (Figure 10). Considered molecular subgraphs can vary by recognizing atom/bond “sequences”, “augmented” atoms and bonds and “atom pairs”.

The sequences are represented by consecutively connected atoms, where types of atom (e.g. C, N, S, etc.) or types of bond (double, triple etc.) or both of them are explicitly shown. Only the shortest distance between the two atoms was used for sequence definition. The Floyd algorithm [74] is used for shortest distance determination. For each type of sequences, the minimal (n_{min}) and maximal (n_{max}) number of atoms in

fragment is defined. For the given combination n_{min} and n_{max} , all intermediate shortest paths with n atoms ($n_{min} < n < n_{max}$) are also counted. In the resulting descriptor vector $D_i(M)$, each locus i is associated to a specific fragment, and its value represents the number of occurrences of that fragment in molecule M .



	Sequences	Atom pairs
Atoms and Bonds	N=C-C-C; N=C-C; C-C-C; N=C; C-C;	N= [4]=C; N= [3]-N; N= [2]=C; C- [3]=C; C- [2]-N;

Figure 10. Example of substructural fragments generated by ISIDA/QSPR software with the fragment size: ($2 < n < 4$)

Since activity varies as the function of the structure [75], machine-learning methods are used to establish the following relationship $P_i = \varphi(D_1, D_2, \dots, D_n)$, where P_i are compounds properties (biological activities or else) of molecules, D_1, D_2, \dots, D_n are molecular descriptors, and φ is mathematical procedure applied to descriptors in order to estimate the property values for the given molecule [76]. In other words, molecular descriptors and compounds activity play the role of independent and dependent variables, respectively.

In this study, Random Forest (RF) and Generative Topographic Mapping (GTM) machine-learning methods were used.

2.1.3 Random Forest

Random forest [77] (either implemented in the CF software [78] or used as R package [79]) was one of methods used in this study. RF is a non-linear machine learning algorithm which is efficient for large databases analysis [78,80,81,82]. This machine-learning method was chosen for this task since it is a non-linear technique which is not of inferior efficiency compared to other non-regression methods.

RF model consists of an ensemble of decision trees built by a Classification and Regression Trees algorithm (CART) [83]. Each tree has been grown according to the following rules:

1. From the whole training set of N compounds a subset of n is sampled using bootstrapping to be used as a training set for one particular tree development. Approximately 33% of the compounds which were not included in the current training set are placed in the out-of-bag (OOB) set. OOB sets are used for cross-validation.

2. A randomly selected subset of m components of the complete, M -dimensional descriptor vector set provides the considered explaining variables. m is tunable and it has a great impact on the models performance.

3. There is no procedure to limit the number of nodes in the tree.

The main features of RF [84] are listed below:

a) there is no need for descriptor pre-selection (descriptor selection is part of the model building process)

b) its non-linear nature supports simultaneous analysis of compounds with different mechanisms of action.

c) the method has its own out-of-bag procedure for the estimation of model quality and its internal predictive ability.

d) models obtained are tolerant to “noise” in source experimental data.

The Applicability Domain (AD) of these QSAR models was calculated using the Euclidean distance-based approach [85]. The distance ($Dist_i$) between the candidate to be predicted and the “center of mass” of the training set in the descriptor space (defined by the mean of all training compound descriptor vectors) was chosen as an indicator of prediction trustworthiness. Compounds for which $Dist_i > Dist_0$ are considered to be outside of the AD. Here, a threshold $Dist_0 = 1.3 \times Dist_{max}$, where $Dist_{max}$ is a maximal distance detected for the training set compounds. The distance between the candidate to be predicted and

the “center of mass” of the training set in the descriptor space (defined by the mean of all training compound descriptor vectors) was chosen as an indicator of prediction trustworthiness: compounds further than a given tunable threshold count as outside of the AD.

2.1.4 Generative Topographic Method

Generative Topographic Mapping (GTM), introduced by Bishop et al [86] is dimensionality reduction technique which transforms the initial, multi-dimensional dataspace into 2D dimensional latent space (also known as GTM map) by fitting a 2-dimensional non-linear manifold into the data space (Figure 11).

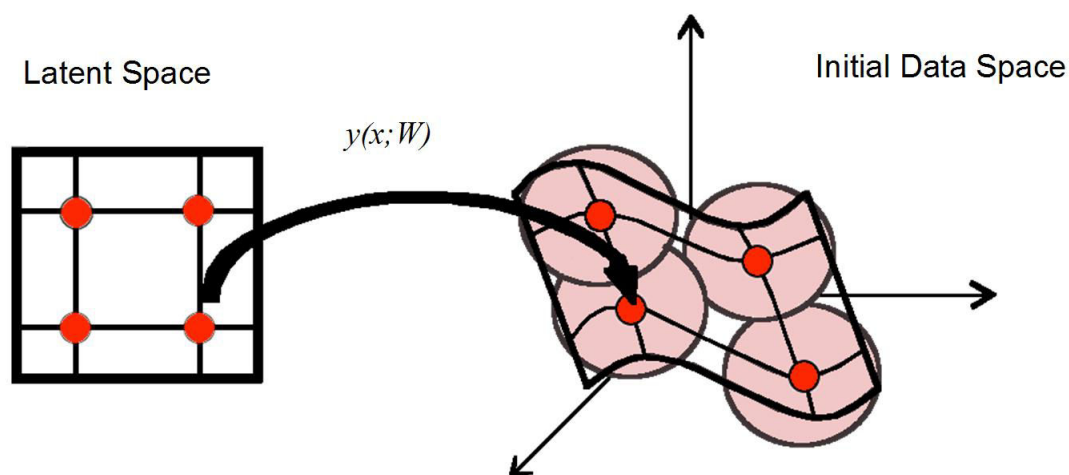


Figure 11. Dimensionality reduction concept. Each node x_k in the latent space (red point on the grid) is mapped to the corresponding manifold point y_k in the initial data space by the non-linear mapping function $y(x; W)$ [87].

The GTM algorithm starts with generation of 2D latent space in the form of a square matrix containing k number of nodes. Each node is mapped to a manifold point y_k embedded in the D -dimensional data space using the non-linear mapping function $y(x; W)$. The manifold points (y_k) are the centers of normal probability distributions (NPDs) of t :

$$p(t|x_k, W, \beta) = \frac{\beta^{D/2}}{2\pi} \exp\left(\frac{\beta}{2} \|y_k - t\|^2\right) \quad (1)$$

where t_n is a data instance and β is the common inverse variance of these distributions. The ensemble of N data instances (in cheminformatics, N molecules) spans the relevant zone of the problem space to be mapped. Molecules are represented by their molecular descriptor vectors t_n , (1...N), which define a “frame” within which the map is positioned, and will therefore be termed “the frame set”.

In Kohonen maps [88] a compound is unambiguously assigned to a node, making compounds within a node indistinguishable. On the contrary, in GTM for every compound projected on the manifold there is a certain probability to “reside” in every node of the grid. The responsibility, or posterior probability, that a point t_n in the data space is generated from the k th node is computed based on current β and \mathbf{W} using Bayes’ theorem:

$$R_{kn} = p(x_n | t_n, W, \beta) = \frac{p(t_n | x_k, W, \beta)p(x_k)}{\sum_{k'} p(t_n | x_{k'}, W, \beta)p(x_{k'})} \quad (2)$$

The responsibilities R_{kn} are used to compute the mean (real value) position of a molecule on the map, $\mathbf{s}(t_n)$ by averaging over all nodes with responsibilities as weighting factors:

$$\mathbf{s}(t_n) = \sum_k x_k R_{kn} \quad (3)$$

Thus, each point on the GTM corresponds to the average position of one molecule. This step completes the mapping by reducing the responsibility vector to a plain set of 2D coordinates, defining the position of the projection point of the initial D -dimensional vector on the map plane. The responsibility vector has the property of being bound to a square grid, a common reference system that may be visually rendered in spite of its still high dimensionality k . A molecule characterized by its r_n vector can be visualized by the pattern of grid nodes that it “highlights”, *i.e.*, with respect to which its responsibility values are significant.

Compounds with nearly identical responsibility vectors are intrinsically related according to the map, and might be thought of as members of a same responsibility-based cluster. It therefore makes sense to use a coarse, binned version of the responsibility vector – the *responsibility pattern RP* – in order to define such responsibility-based clusters as compounds sharing a same RP. The “binning” process of real-value r_n to integer RP_n is done as follows: If the responsibility of molecule n for node k is below

what is empirically considered “below the minimally relevant threshold” – empirically established at 1%, the corresponding integer responsibility level is set to zero. Beyond this threshold of 0.01, any additional 0.1 units of responsibility contribute an increment of +1 to the R_{nk} value, *i.e.* $R_{nk} = 1$ if $0.01 \leq R_{nk} < 0.11$, $R_{nk} = 2$ if $0.11 \leq R_{nk} < 0.21$, *etc.* [89]. Formally, one may therefore define:

$$RP_{nk} = [10 * r_{nk} + 0.9] \quad (4)$$

where the $[..]$ operator means truncation. If molecules are members of the same responsibility-based clusters, they must be structurally similar.

GTM can be used as a classification tool. In this study a Latent-Space Classification approach was used [71,72], as outlined in the following. Given a training set of m molecules assigned, on the basis of experimental input, to different and non-overlapping categories c_i (typically, actives $\in c_1$, inactives $\in c_2$), then the responsibility vector of each molecule can be used to transfer class information onto its associated nodes [71,72]. Intuitively, if the class assignment is visualized as a color, then each molecule will “transfer” some of its color to the nodes, proportionally to responsibilities. Transferred colors accumulate in the nodes, eventually defining nodes where one specific color dominates over the others (Figure 12).

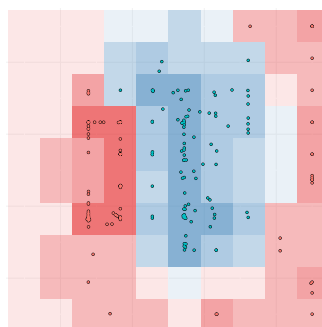


Figure 12. Example of GTM latent space classification model with applicability domain (AD) [71,90] for DUD [91] AchE inhibitors (red) and decoys (blue). Lighter regions have a lower probability of association to the winning class $P(x_k|C_{best})$ and may therefore be discarded from the applicability domain of the model. The points on the map represent individual compounds colored by class [87].

Mathematically, the (normalized) amount of color on each node represents the probability of association of the node k to class c_i :

$$P(c_i|x_k) = \frac{P(x_k|c_i)P(c_i)}{\sum_j P(x_k|c_j) P(c_j)} \quad (5)$$

where $P(x_k|c_i)$ is computed as follows:

$$P(x_k|c_i) = \frac{\sum_{n_i}^{N_{c_i}} R_{kn_i}}{N_{c_i}} \quad (6)$$

where R_{kn_i} responsibility of node k for a molecule belonging to class c_i , n_i enumerates training set compounds belonging to class c_i , N_{c_i} is the number of training set compounds belonging to class c_i , and $P(c_i) = \frac{N_{c_i}}{N}$, represents the prior probability of class i , *i.e.*, the fraction of class members within the training set.

If $P(c_1|x_k) > P(c_2|x_k)$, node k will be formally assigned to class 1, and visually rendered in the associated color (Figure 12), with an intensity modulated by $P(x_k|c_i)$. This allows checking whether the local dominance of class 1 corresponds, indeed, to a significant local accumulation of members of that class, or whether the prevalence is the result of unreliable extrapolations of distribution tails to nodes far off the actual regions of interest.

Now, “colored” nodes represent a repository of the knowledge extracted from the training set compounds, and can be subsequently used for predictions, by transferring the acquired “color” back to query compounds q to be classified. As a first step, a query compound q defined by its descriptor vector t_q will be located on the GTM, *i.e.*, associated to responsibilities $\{R_{kq}\}$, and optionally mapped to its 2D residence point s . In this study, the so-called local method was chosen for definition of projected compounds class. The local method based on the 2D representation only uses the conditional probability of the node closest to the molecule in 2D, $P(x_{nearest}|c_i)$:

$$P(c_i|t_q) = P(x_{nearest}|c_i) \quad (7)$$

The local method was chosen by the evolutionary procedure used for map building (*vide infra*) out of other possible options, as the one yielding optimal cross-validation results (map fitness). This approach is also the most intuitive one, as it allows direct reading of molecular properties from (latitude, longitude) specifications. In order to translate $P(c_i|t_q)$ into a clear-cut answer to the question “to what class does q belong”, it is sufficient to consider the largest of these values as “winning” class, although the

confidence in the prediction should be downgraded if the winning class won by a narrow margin only [92].

Studies dedicated to GTM modeling highlighted the fundamental distinction between actual unsupervised map (manifold) construction, based on a frame set, and subsequent (supervised) learning or “coloring” of this map, based on a potentially different training set. Some options or parameters only concern the unsupervised manifold fitting step, and include the four GTM setup parameters: the grid size k , the number of RBFs M , the RBF width factor (w), and the weight regularization coefficient (λ), in addition to the frame set choice, which can be formally regarded as an additional degree of freedom. Eventually, one meta-parameter of paramount importance affects both manifold construction and learning process: the choice of the initial descriptor space, the primary conveyor of numerically encoded structural information. All these parameters have an impact on the quality of the final predictive model supported by the manifold. If a map is designed to describe the chemical space of compounds possessing a certain property, map quality must be evaluated by its classification capacity. Thus, an evolutionary algorithm needed to choose the best among models based on the same frame set but different parameters and descriptors can be used. Choices of parameters and descriptors can be synthetically represented as a “chromosome”, with loci dedicated to each mentioned degree of freedom. Some loci represent categorical variables, denominating the choice of frame set, descriptor type or prediction method; some are integers (size, RBF number), and others are real numbers. Evolutionary computing readily supports browsing such heterogeneous search spaces, which makes it a method of choice for the quest of optimally tuned GTM models. The chromosome (“genotype”) unambiguously encodes the “recipe” to build a GTM model (the associated “phenotype”). This phenotype is defined by the ability to “survive” in the competitive environment of a fixed-size chromosome population (under steady evolution through crossover and mutation events involving current members), *e.g.*, its “fitness” score. The nature of this fitness score has already been hinted at: some mean of cross-validated predictive power scores, over selection sets. This might be refined by introducing a penalty related to the spread (standard deviation) of individual scores per set: at equal mean predictive power, the map performing roughly equally well for each selection model is to be preferred to a map doing very well on few models but failing for others (Figure 13).

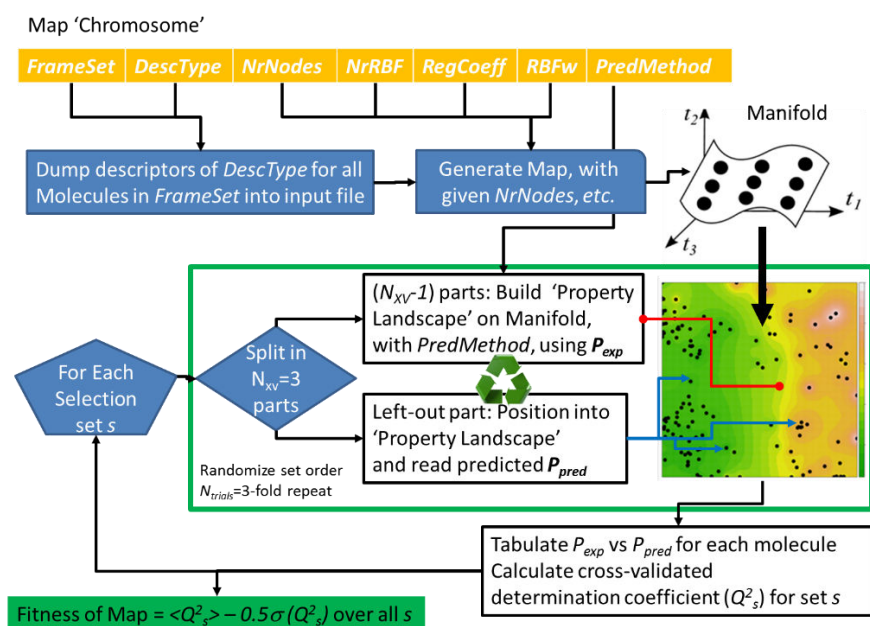


Figure 13. Scheme of the detailed process of estimating the fitness score for a multiproperty-competent GTM model operating in regression mode, and employing repeated, randomized leave-1/3-out cross-validation for a robust assessment of individual quality criteria Q^2 for each selection set [87]

2.1.5 Statistics used for QSAR models performance assessment

For regression models, the predictive ability is estimated by root mean-squared error (RMSE) and coefficient of determination (R^2):

$$RMSE = \sqrt{\frac{\sum_{i=1}^n (y_{pred,i} - y_{obs,i})^2}{n - 1}} \quad (8)$$

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_{pred,i} - y_{obs,i})^2}{\sum_{i=1}^n (y_{mean} - y_{obs,i})^2} \quad (9)$$

where n is the number of compounds in a test set; $y_{obs,i}$ is an observed activity value of i -th compound in a test set; $y_{pred,i}$ is predicted activity value of i -th compounds in a test set; y_{mean} is a mean activity value for compounds of a training set.

As for classification models, there are also several statistical parameters for reliable classification performance evaluation. The Confusion matrix, given in **Table 1**, presents all outcomes for 2-class model prediction:

Table 1. A confusion matrix for 2-class classification, where True Positives – Active compounds predicted as active, True Negatives – Inactive (or activity unknown) compounds predicted as inactive, False Positives - Inactive (or activity unknown) compounds predicted as active, False Negatives - Active compounds predicted as inactive

Class/Predicted	as Active	as Inactive
Active	True Positives	False Positives
Inactive	False Negatives	True Negatives

Among the various evaluation criteria, the statistical measurements which were used in the current work are precision, recall, sensitivity, specificity, balanced accuracy and Cohen's kappa coefficient (κ).

Precision is defined as the ratio between true positives and all the positives

$$Precision = \frac{TP}{TP + FP} \quad (10)$$

Sensitivity (or Recall) is the proportion of correctly identified positives in the set of all positives

$$Sensitivity (Recall) = \frac{TP}{TP + FN} \quad (11)$$

Specificity is the proportion of negatives which are correctly identified as such.

$$Specificity = \frac{TN}{TN + FP} \quad (12)$$

Balanced accuracy assesses the overall predictive capacity of the classifier

$$Balanced\ accuracy = \frac{Sensitivity + Specificity}{2} \quad (13)$$

Cohen's kappa coefficient is also used for classifier performance assessment if the dataset is small and compounds are distributed in classes disproportionally

$$\kappa = \frac{p_o - p_e}{1 - p_e} \quad (14)$$

where p_o is

$$p_o = \frac{TP + TN}{\text{all predictions}} \quad (15)$$

and p_e is

$$p_e = \left(\frac{(TP + FN) \times (TP + FP)}{\text{all observations}} + \frac{(TN + FN) \times (TN + FP)}{\text{all observations}} \right) / \text{all observations} \quad (16)$$

where TP – True Positives, FP – False Positives, TN – True Negatives, FN – False Negatives

Cohen's kappa computes the ratio between the chance-corrected agreement of the accuracy in the numerator and the chance-corrected perfect agreement in the denominator [93]. This ratio yields an estimate of how much better the actual agreement is over chance agreement. The values for kappa range between -1 and 1: a perfect model gives a kappa value of 1, whereas kappa values lower than 0 indicate models performing worse than random. Model prediction is considered substantially different from random if $\kappa \geq 0.21$. Cohen's kappa was used in this study in order to assess non-randomness of prediction for models based on small training set.

Model robustness and its predictive power are assessed by cross-validation and external testing, respectively. In this study the robustness was estimated either by a k-fold cross-validation procedure [94] or the out-of-bag technique described earlier. In k-fold, the whole dataset is split in **k** non-overlapping pairs of training and test sets. Each training set covers $\frac{k-1}{k}$ of the data set, and corresponding test set is composed of the remaining $\frac{1}{k}$. This ensures an external prediction for every molecule from the modeling set. External test set consists of compounds which were never used in model build and, therefore, present a challenge in terms of activity prediction based on previously not used structural information.

2.1.6 Pharmacophore modeling

A pharmacophore model is an ensemble of steric and electronic features of the ligand, providing specific ligand-biological target interactions responsible for triggering (or blocking) biological response [95]. Broadly used pharmacophore features include H-bond acceptors and donors, charged or polarizable groups, hydrophobic fragments and aromatic rings. The use of these features expands the concept of bioisosterism [96],

which recognizes that certain changes in structure of biologically active compounds does not result in activity disappearance due to similarity of certain substituents contribution to activity.

The three-dimensional pharmacophore model does not only include the above-mentioned features but also specifies the Euclidian distance between all of them.

There are two major strategies for pharmacophore model development: using 3D structure information from ligand-target complexes (structure-based modeling) or using information on active compounds structure only (ligand-based modeling) [96].

Compounds used for virtual screening must be represented by a set of conformers, amongst which some must correspond to spatial restrictions posed by pharmacophore model in order to be found active. Compounds with conformers matching a user-specified number of model features form a hit list [97]. The molecules ranking within the hit list, as well as the degree of pharmacophore model matching is determined by a scoring function. [98]

LigandScout was used for pharmacophore modeling in this study. This software differs from other packages (Catalyst, MOE and Phase) in terms of alignment algorithm efficiency. In this algorithm, the first step is the generation of the 3D pharmacophore features identified for each training set compound conformer. Next is calculation of inter-feature distances for each feature type. A pairwise comparison of distance sets calculated for the pharmacophore model and for the conformer pharmacophore features is taking place afterwards. Pair assignment is performed using the so-called Hungarian matching algorithm and the feature distances minimization between model and compounds conformer using Kabsch alignment algorithm was carried out. [99,100].

Alignment quality was estimated via four different in-built LigandScout scoring functions. The pharmacophore fit score is a geometric scoring function. It favors solutions with a high number of geometric matched feature pairs, while penalizing ones with higher Root-mean-square-deviations (RMSD) between model and conformer feature. Atom overlap score is defined by overlap of atom van der Waals spheres, whereas Gaussian function representation of molecular volume overlap is measured to calculate Gaussian shape similarity score. The fourth scoring function is a combo score of the first two scores and named pharmacophore fit and atom overlap score [101]. In this study the simplest pharmacophore fit scoring function score was used (see eqs. 10 and 11).

$$S_{RMS} = 9 - 3 \times \min(RMS_{FP}, 3) \quad (17)$$

$$S_{FCR} = c \times N_{MFP} + S_{RMS} \quad (18)$$

where S_{RMS} is RMSD score the matched feature pair in the range varying from 0 (no match at all) to 9 (perfect match)

RMS_{FP} is the matched feature pair distances RMSD

S_{FCR} function for alignment quality assessment

c is a weighting factor for multiplying matched feature pairs (currently 10.0)

N_{MFP} - the number of all matched feature pairs.

There are two approaches for ligand-based pharmacophore model generation. Model generation is a pairwise process, meaning that at each step one pharmacophore for two compounds is created. This is achieved by selecting common features of training set compounds (Shared feature pharmacophore) or by augmenting all features of a training set (Merged feature pharmacophore). In the second case, each feature is scored and those that do not match all input molecules are removed.

In LigandScout the ligand-based approach allows clustering the ligands to simplify the search for similar patterns of interactions with a target macromolecule. After generating conformers for all compounds, they are clustered according to the RMSD values calculated between centers of corresponding pharmacophores for a pair of conformers of selected compounds.

In order to validate the model, a set of both active and decoy compounds is used. Decoys represent molecules similar to those in the training set, although the main the desired activity can be absent [102]. Decoys are usually selected from some random small-molecule compounds database, while the validation set of active compounds is usually a set of actives not used in model development, as in QSAR model build. The validated model is ready to be used in the virtual screening and several pharmacophore models can be used simultaneously.

In the LigandScout the "conventional" pharmacophore features shown in Figure 14 are implemented.

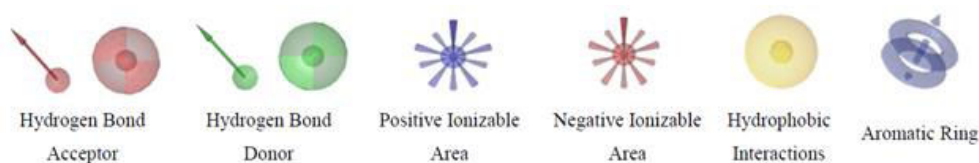


Figure 14. "Conventional" pharmacophore features [101]

2.1.7 Third-party predictive models used in Virtual Screening

PASS (Prediction of Activity Spectra for Substances) was used for assessment of screened compounds potential polypharmacology, toxicity or adverse effects. Predicted activities include mechanisms of action (5-HT, GABA_A inhibition, etc.), pharmacological effects (e.g. anxiolytic, antiemetic, aphrodisiac, etc.), specific toxicities (mutagenicity, fetotoxicity, teratogenicity, etc.) and metabolizing paths enzymes (CYP2C9 substrate, CYP3A4 substrate, etc.) [103]. The PASS algorithm is based on the structure-activity relationship analysis (SAR) for the training set of more than 60000, marketed drugs, drug-candidates, leads and toxicants with experimentally determined activities. Activity predictions are given as a list of activity types, with the probability of presence (P_a) and absence (P_i) for each particular activity. By default, $P_a > P_i$ value was used as a threshold that provides the mean accuracy of prediction about 90 % in leave-one-out cross-validation for training set. However, the user can define a threshold P_a value according to his own conception of plausible activity occurrence.

2.2 Databases

In this study, several small-molecule databases, such as BioinfoDB, PubChem, PCI, Zinc and ChEMBL, were used for virtual screening and chemical space analysis.

BioinfoDB [104] is a database of commercially available compounds. The 14.1 version comprising 3 207 317 compounds was used.

PCIdb is a combinatorial library of virtual compounds from Physico-Chemical Institute, Odessa. This database consists of 288 structurally similar virtual compounds generated as a combination of scaffolds and some typical fragments from previously synthesized compounds with a DNA affinity potential. Both BioinfoDB and PCI were used as a source of antiviral candidates in the virtual screening.

PubChem [105] contains information on roughly 220 mln substances from approximately 400 sources like Chemical vendors (e. g. Enamine) and Research and Development Institution (e. g. Southern Research Institute). It was used to check the novelty of screened compounds.

Zinc [106] version 12 with 35 mln purchasable compounds from over 100 vendors (e. g. Aldrich CPR) was used as a source of decoys for pharmacophore modeling.

ChEMBL [107] version 19 (July 2014) comprising approximately 1,4 mln compounds and activity data from more than 1mln bioassays was used as a source of data for antiviral chemical space analysis.

2.3 Data curation tool

The Konstanz Information Miner (KNIME) [108] is a modular environment, which provides comprehensive visual assembly and interactive execution of a data pipeline. It ensures data processing and visualization in the shape of interconnected modules or nodes.

This environment allows constructing and adapting the analysis flow using standardized building blocks, which are then connected through data or models transferring pipes (Figure 15). The advantage of this system is the intuitive, graphical way to record the workflow.

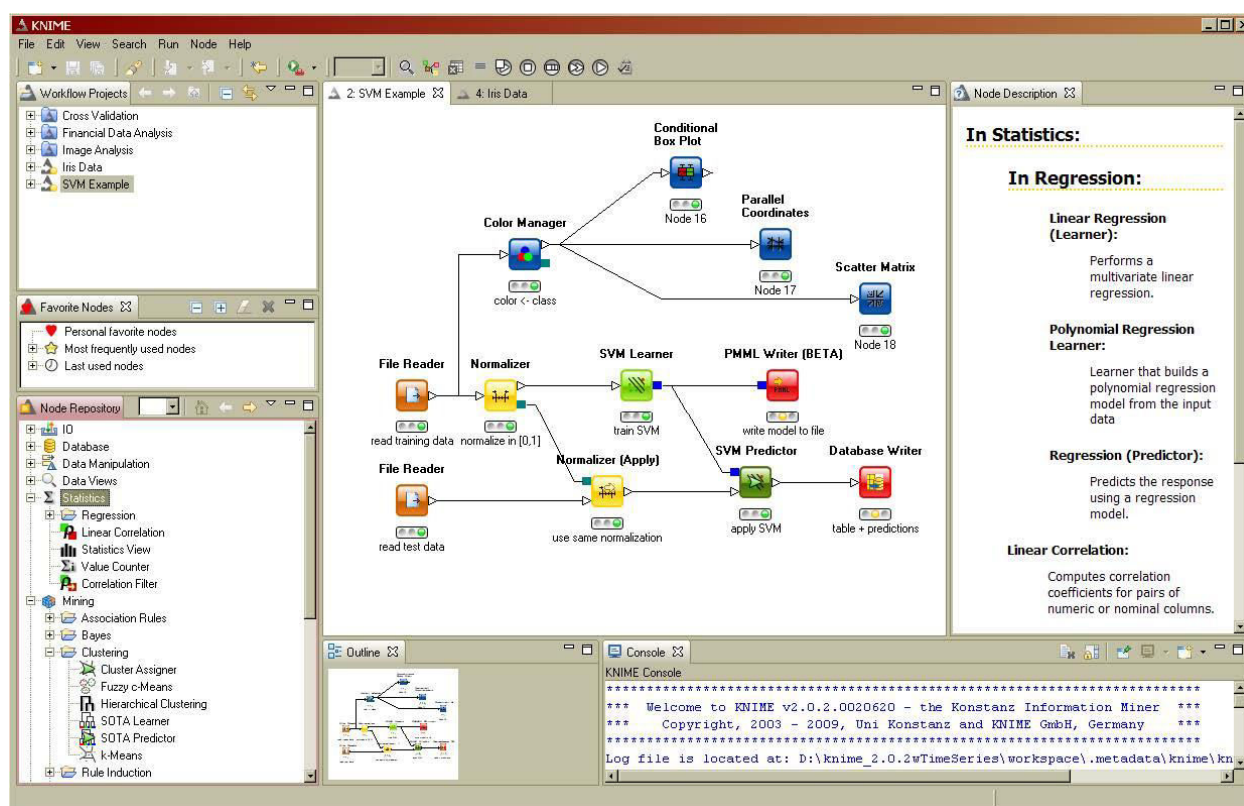


Figure 15. KNIME “workbench” example [108]

Nodes in KNIME are the most general processing units. Each node has a pre-defined import and export instances, for data or models transport. The software provides a large variety of nodes, one for data sources selection, data preprocessing steps, model building algorithms, as well as visualization tools. Nodes are dragged onto the workbench, where they can be connected to previously used ones.

A flow typically starts with a node that imports data from a certain source, such as text files or databases. Extracted data is stored in a Knime-specific table-based format consisting of columns with a certain data type (integer, string, image, logical, etc.) and a casual number of rows corresponding to the column content. These data tables are transferred to other nodes that modify, process, model or visualize the data. Modifications embrace handling of missing values, filtering by column or row values, oversampling, merging and dividing tables etc. After data preparation, the development of predictive models using machine learning or data mining algorithms, such as decision trees, regression equations or support vector machines can be built. Several view nodes are available for the visualization of analysis results, whether it is the processed data or developed models.

KNIME offers a large variety of nodes, comprising the ones for various types of data import, export, manipulation, and modification, as well as the most commonly used data mining and machine learning algorithms and a number of visualization components. Another type of nodes is wrappers, which integrate functionality from third party libraries. In particular, KNIME integrates functionality of several open source projects that cover major areas of data analysis such as Weka [109] for machine learning and data mining, the R environment [110] for statistical computations and graphics, and JFreeChart [111] for visualization. One of the important design decisions was to allow users to modify the workflow easily, namely adding new nodes and data types.

Workflows in KNIME are in the nutshell graphs connecting nodes, namely a direct acyclic graph (DAG). The workflow manager allows the inclusion of new nodes and addition of directed edges (connections) between two nodes. It also keeps track of the status of nodes (configured, executed, ...) and gives back, on demand, a pool of executable nodes. This way the environment framework can freely distribute the workload among a couple of parallel threads or even a cluster of computer servers.

Unlike some workflow or pipelining tools, nodes in KNIME process the entire input table before the results are sent to consequent nodes. This allows each node to keep its

results permanently and therefore, workflow execution can be easily stopped at any node and resumed afterwards. Intermediate results can be viewed any time and new nodes can be included in the form of created blocks without preceding nodes having to be re-executed. The data tables are kept together with the workflow structure and the nodes' settings.

The integration of these and other tools not only enriches the functionality available in KNIME but has also proven to be helpful to overcome compatibility limitations when the aim is on using these different libraries in a shared setup.

2.4 Scaffold analysis tool

A scaffold is a (poly)cyclic molecular core framework which can be seen as the quintessential feature defining the compound class emerging when adding substituents to it. Even though scaffolds can be derived from a dataset by visual inspection, nowadays powerful computational tools exist for accomplishment of this task. The "Scaffold Hunter" [112] software was used in this study. This program is using a scaffold network algorithm, which is a modification of a scaffold tree [113]. The algorithm provides a classification of chemical scaffolds which form the leaf nodes (molecular framework) in the hierarchy trees. By an iterative removal of cyclic fragments, scaffolds which form the higher levels in the hierarchy tree are obtained. Prioritization rules ensure that less significant, peripheral rings are excluded first. All scaffolds in the hierarchy tree are clearly defined as meaningful chemical entities making the classification chemically intuitive. The classification procedure does not depend on dataset composition and scales linearly with the number of compounds.

Two scaffolds are regrouped into a same node if the summed scores of the transformations needed to turn one into another according to the proposed scoring scheme does not exceed a given threshold. Since each scaffold in the classification tree has only one parent scaffold, it is important to select the prioritization rules carefully in order to keep that part of the scaffold as a parent which characterizes it in a chemically intuitive way.

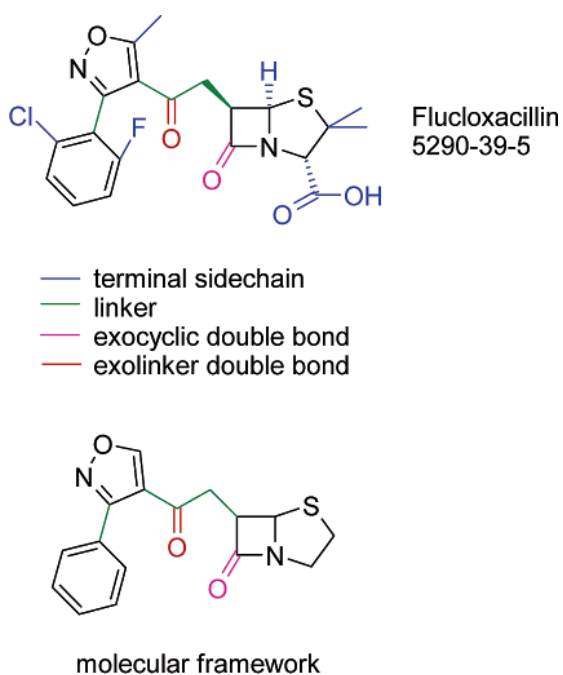


Figure 16. Example of obtaining primary molecular framework [113]

The classification begins by removing all terminal side chains to determine primary molecular framework. Exocyclic double bonds, and double bonds directly attached to the linker (“exolinker double bonds”) are kept (Figure 16). It is done to ensure that planar sp^2 carbon atoms are recognizable in the scaffold and are not converted into tetrahedral sp^3 carbon atoms which would lead to unwanted changes in fragment geometry.

The stereochemistry is discarded at the stage of molecular framework determination. Even though retaining the information about the stereocenters presence in scaffold is useful, scaffold tree algorithm discards it due to unavailability of stereo information for all compounds in many databases. Partial or incomplete data on 3D structures within the dataset can lead to errors as the outcome of the classification would depend on. Considering the fact that in SAR tasks 2D descriptors performance is as good as 3D descriptors [113], the loss of stereochemistry information can be expected to have little impact on scaffold space description as well.

After sidechains disposal, rings are removed iteratively one by one until only one cyclic fragment remains. Removal of a ring means that bonds and atoms which are part of the ring are removed, unless atoms and bonds belong to any other ring. In addition, all exocyclic double bonds attached to the atoms of removed ring are discarded as well. If the removed ring is connected to the resulting scaffold by an acyclic linker, this linker is considered a terminal side chain and is removed as well. If the ring removal would lead to a disconnected structure, this ring cannot be removed.

The scaffold network method used in Scaffold Hunter is an advanced version of the scaffold tree algorithm. It explores all branches rather than picking a specific scaffold at each hierarchy level. For the three 5-HT₃ antagonists in Figure 18, the scaffold network approach created the green scaffolds in addition to the blue scaffolds that were generated by the scaffold tree approach.

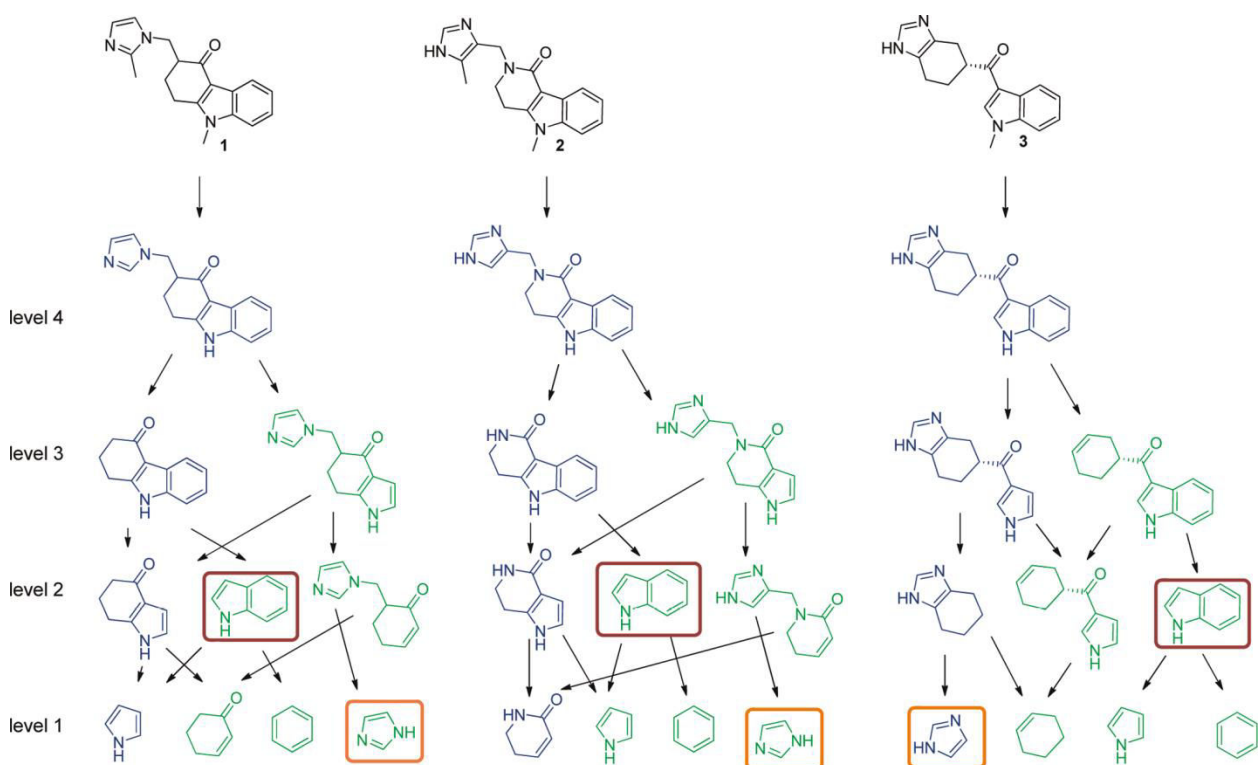


Figure 18. Example of scaffold network structure [112]. Note that the same scaffold, such as indoles (red outline) and imidazoles (orange outline) can be derived from previously distinctive tree branches using the improved methodology.

PART 3 QSPR MODEL FOR AQUEOUS SOLUBILITY PREDICTION

Building a QSPR model capable of predicting solubility in water for structurally diverse set of antiviral compounds was needed for the virtual screening described in the Part 4. The main originality of the herein advocated approach is to explicitly include the temperature dependence of solubility into this therefore original structure-property model. In this part of the thesis, solubility dataset preparation, QSPR model development and validation is described.

3.1 Dataset preparation

The data on aqueous solubility of a large set of compounds at different temperatures was taken from Handbook of Aqueous Solubility Data [114]. However, not all compounds listed in the handbook were used in model development. To perform the assessment of data accuracy, we applied the data evaluation system presented in [114] (Table 2). It consists of 5 parameters (Temperature, Purity of solute, Equilibrium time/agitation, Analysis, Accuracy and/or precision) evaluated using 3 grades: 0,1,2 (low, medium, high). Only compounds, which had Temperature, Purity of solute and Accuracy, and/or precision criteria assessed at least as medium, were chosen for further study. Secondly, some classes of organic compounds (namely organic salts, polymeric compounds and crystalline hydrates) were excluded from data set due to difficulties of their representation by molecular descriptors. Also, it was crucial to remove mixtures, duplicates and compounds with ambiguous CAS number.

Table 2 Explanation of Evaluation Scores from solubility data source. ^a Parameter acronym.

			Score	
P ^a		0	1	2
T	Temperature	Not given, ambient, or room temp	Given with no range	Given with range
P	Purity of solute	Not stated or as received	Stated with no range or as received	Stated with range or altered with range or calculated
E	Equilibration time/ agitation	Not stated	Stated briefly	Described in detail
A	Analysis	Not stated	Stated briefly or stated in other paper	Described in detail
A	Accuracy and/or precision	1 significant figure or range > 20%	2 significant figures or range 5–20%	3 significant figures or range 1–5%

As a result, 1484 aqueous solubility data points in the temperature range 4-97 °C for 562 organic compounds have been selected. They cover various classes used in medicine (e. g. barbituric acid, benzodiazepines derivatives), agriculture (e. g. thiophosphate pesticides), and military (e. g. nitroaromatics). Solubility was expressed in mol/L, with data points dispersed between -11.9 and 1.18 log units (logSw). Among members of the data set, 141 compounds have solubility data for at least 3 temperatures which allowed determining solubility-temperature curves needed for solubility temperature coefficient determination.

3.2 Model development

Solubility model development consists of 5 stages (Figure 19).

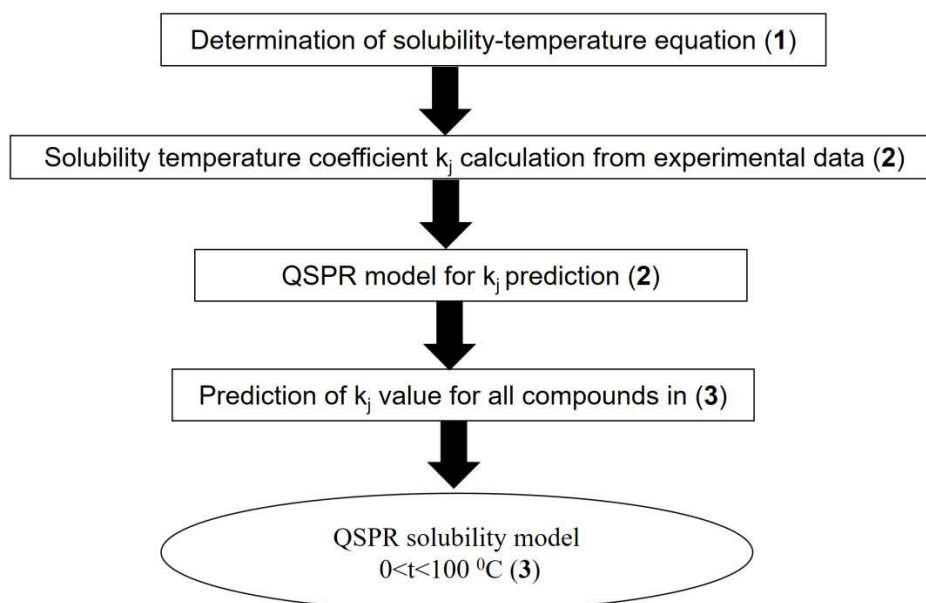


Figure 19 Workflow of the solubility model development. Numbers in brackets refer to sets in Figure 20

3.2.1 Determination of solubility-temperature equation

Solution can be viewed as a system of two components (solute and solvent) with an equilibrium between two-phase and homogenous system. Therefore, the application of Van't Hoff's equation (19) to the process of dissolution [115] results in the following:

$$\ln x = \frac{\Delta H_{fus}}{R (T_m - T)} \quad (19)$$

Where, x is the mole fraction of solubility, T_m is the temperature of melting point, T is the temperature of dissolution and ΔH_{fus} is the enthalpy of fusion. However, the use of this equation for solubility prediction is limited since it is correct mostly for solutions with the solid solute and enthalpies of fusion are not always available. This means a simpler equation to describe temperature-solubility relationship is needed. To achieve this goal, initially, we selected a subset (#1 in Figure 20) from the training set (#3 in Figure 20). These are compounds with solubility data points for which multiple solubility measures at several distinct temperatures were determined. We consider these compounds to be representative with respect to the training set, since their solubility ranges from highly soluble ($\log Sw = 0.9$) to practically insoluble substances ($\log Sw = -11.89$) and they are both solids and liquids with molecular mass varying from 84 (Dicyanodiamide) to 499

(2,2',3,3',4,4',5,5',6,6'-Decachlorobiphenyl). Moreover, there are different combinations of size, solubility and state among them, such as poorly soluble small liquids (3,3-Dimethylpentane), poorly soluble big solids (Decachlorobiphenyl), highly soluble small solids (Succinic acid) etc.

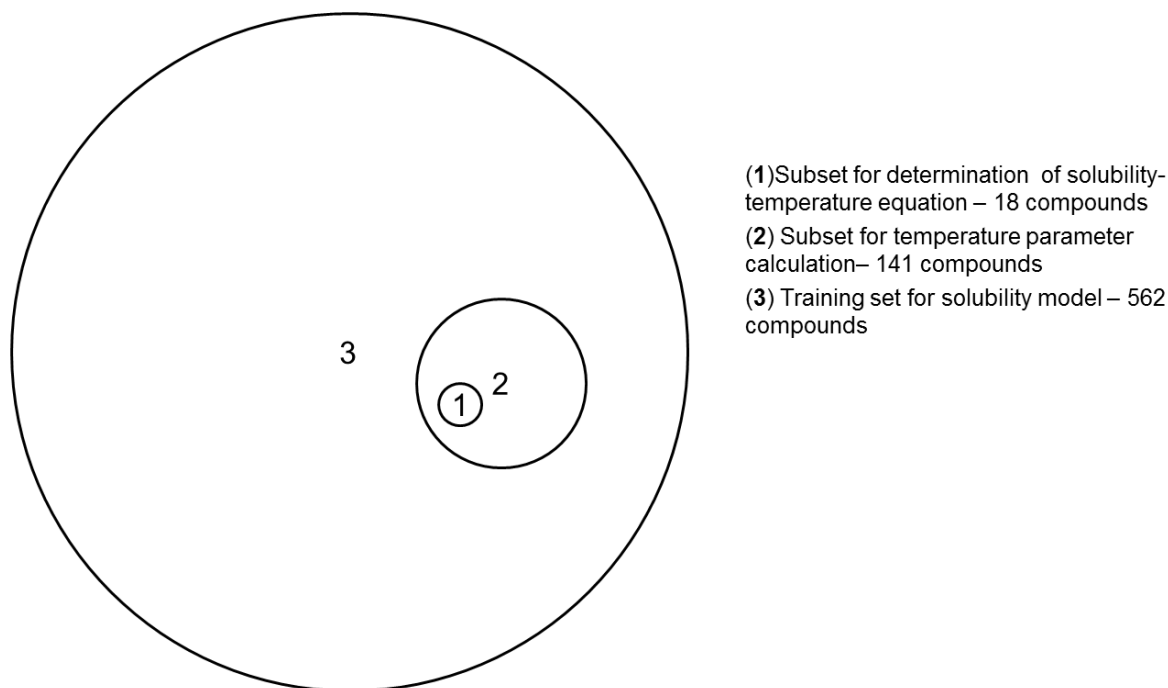


Figure 20 Description of sets used in solubility model development

To define the empirical equation which will be used further in our QSPR study, TableCurve 2D software [116] was used. Equation fit was determined by F-test values as follows: for every compound top 10 equations with the highest F-test values were selected, then equations were ranked according to their occurrence within data set. The most “fitting” equation was linear one (20), since it was present in top 10 list of 16 out of 18 compounds. Appendix A Table S1 collects the F-test values for the best equations.

$$\log Sw_j = k_j T + c_j \quad (20)$$

In eq.(20), k_j is a coefficient of j th compound and c_j is j th compound solubility at $T = 0^\circ\text{C}$. Linear equation (20) has the simplest form and it indicates that usually temperature rise will lead to increase in solubility. Therefore, this equation was used for QSPR modeling in our study.

3.2.2 QSPR model for k_j prediction

Solubility data from Subset #2 were used to derive k_j for each of 141 compounds with the help of Microsoft Excel tool. Obtained values formed a training set (Appendix A **Table S2**) to build a model for k_j as a function of chemical structure. The SiRMS descriptors and RF we used for the model development. Both out-of-bag procedure and 5-fold cross-validation were applied to evaluate model's robustness. Resulting model had R^2 and RMSE of 0.75 and 0.034 for oob & 0.78 and 0.066 for XV, respectively. Considering the fact that k_j has a rather narrow distribution, models characteristics look acceptable. The developed model has been used to calculate k_j for all compounds from the set #3.

3.2.3 QSPR solubility model development

All solubility data in set #3 were used for the model building. Apart SiRMS descriptors, the k_jT value was used as an additional descriptor [117]. Resulting RF model performs pretty well: $R^2 = 0.96$ and 94 and RMSE = 0.21 and 0.38 for oob and XV, respectively. This RMSE value is comparable the experimental error of solubility measurements estimated as 0.24 log units. [118]

3.3 Model validation on external test set

Even though cross-validation is a powerful tool for evaluating model's quality, it was decided to use an external test set for model validation as well. Therefore, 5 compounds (42 data points) which were not used in previous training and test sets with solubility within 5-81 °C temperature range obtained from different sources [119,120,121,122,123] were selected for this purpose.

For these 5 compounds, the comparison between experimental data and QSPR predicted solubility values shows fairly acceptable RMSE = 0.77.

The last important test for our solubility model was comparison with another computational approach's predictive performance. For this purpose, we have selected the results of our recent study [57] where we predicted the temperature dependence of solubility for nitro-compounds within COSMO-RS [124] approach. Since six compounds investigated in [57] are already included in our model's training set, we have selected four remaining compounds that have in total 18 solubility data points for comparison. The results of such comparison are presented in the Figure 21.

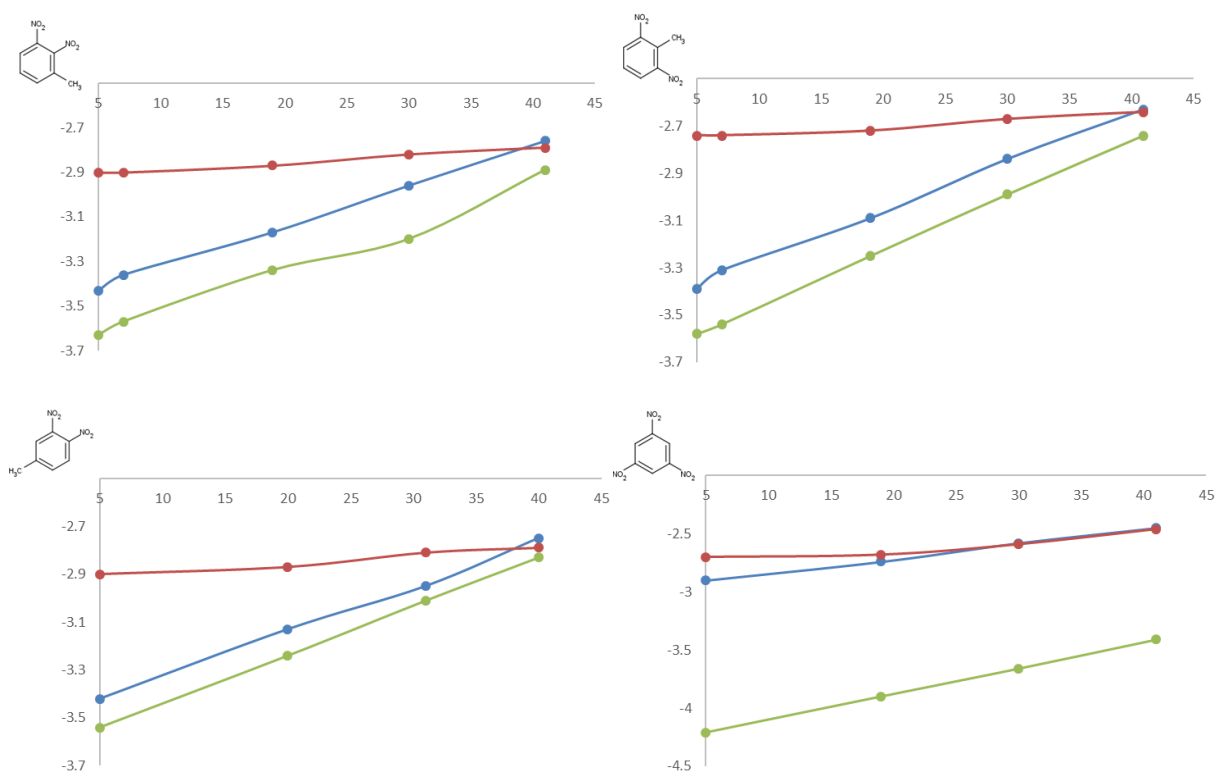


Figure 21. Comparison of COSMO-RS and our model predictive performance. X-axis is temperature, Y-axis is the logSw, blue – observed solubility, red – our models prediction, green – COSMO prediction.

Since Random Forest model has slightly better accuracy in predicting solubility of above-mentioned compounds compared to COSMO-RS approach (see RMSE comparison) data placed in Appendix A Table S3), we expect that it is slightly more accurate. Moreover, solubility values for these compounds were calculated within several seconds what is not possible using quantum chemical calculations. Also, in contrast to COSMO-RS data, the developed QSPR model shows the pattern of solubility similar to the experimental data. To illustrate this, we present the patterns of the solubility at 30 °C (Table 3)

Table 3 Comparison of selected nitroaromatic compounds solubility at 30 °C

	Compounds ordered according to logSw value
COSMO-RS	
QSPR Predictions	
Experimental Data:	

However, to be more conclusive in the comparison of the performance of Random Forest and COSMO-RS methodologies one needs to obtain similar temperature dependence values of water solubility for the same number of compounds that has been considered for Random Forest level. This is out of the scope of this particular work.

Solubility models described here were used in virtual screening stage of CADD. According to the model, 100 compounds were soluble enough and 68 of them were chosen for biological testing. Even though bioassay required compounds to be soluble in 1:4 DMSO-water solution instead of 100% water, only 19 compounds turned out to be insoluble. Which means solubility prediction for screened compounds was correct in 72% of cases.

3.4 Conclusions

We determined that the value of temperature of dissolution (T) by itself is not the best option of a descriptor that could be used in QSPR analysis to predict a temperature dependence of water solubility. Such a descriptor is the product between regression coefficient k of equation (26) and the temperature of dissolution. Based upon this analysis we have developed two step QSPR procedure to predict temperature dependence of water solubility of organic compounds. The first step uses SiRMS generated descriptors to predict the value of k_jT . The second step applies both SiRMS generated descriptors and a value of kT , to generate effective models that are able to accurately predict the

temperature dependence of solubility. The successful predictive ability of these models has been illustrated by the application of independent external test set and the comparison with limited amount of the temperature dependent water solubility values for the compounds obtained at COSMO-RS level.

PART 4 COMPUTER-AIDED DESIGN OF BROAD-SPECTRUM ANTIVIRALS

In this study, virtual screening tools were designed to select nucleic acid intercalators. Nucleic acid intercalation is considered to be the targeted property which ensures broad-spectrum antiviral activity, since intercalation distorts nucleic acids conformation, hindering viral reproduction. This mechanism does not depend on viral protein composition, and thus should not be affected by viral mutations. This study was carried out as a part of a larger antiviral research project at A.V. Bogatsky Physico-Chemical Institute NAS of Ukraine.

The design of antiviral compounds consists of three sections:

- Datasets preparation for the modeling and virtual screening (4.1)
- models development and validation (4.2)
- virtual screening of selected databases (4.3)

4.1 Data preparation

167 DNA intercalating compounds with associated antiviral activity data were used for both pharmacophore and QSAR model development (see Appendix B Table S4). Each molecule contains a polycyclic planar fragment linked to basic amino group to provide a stacking interaction with the nucleic acid. According to the type of polycyclic fragment, the dataset consists of seven classes, as it is shown in Figure 23.

DNA affinity measurement described in [125] was used to define binding constants (K_i) for all 167 compounds. Thereof, 161 compounds with $\lg(K_i) \geq 4$ have been recognized as reasonable DNA intercalators and used for pharmacophore model development. The maximum antiviral effect E_{\max} (%) within 0.2 - 620 μM concentration range (Figure 22) was measured as described in [8] for 117 compounds (see Appendix B).

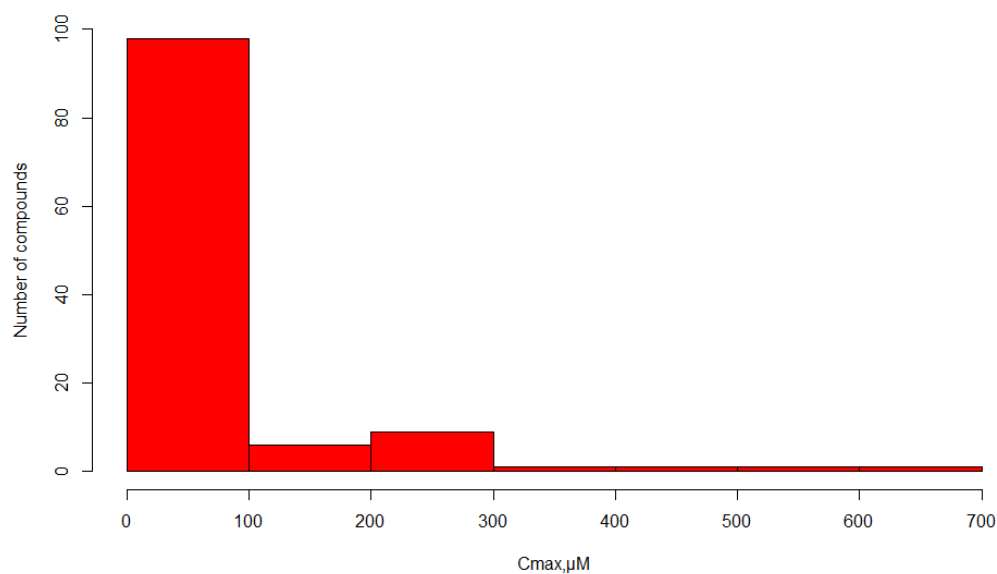
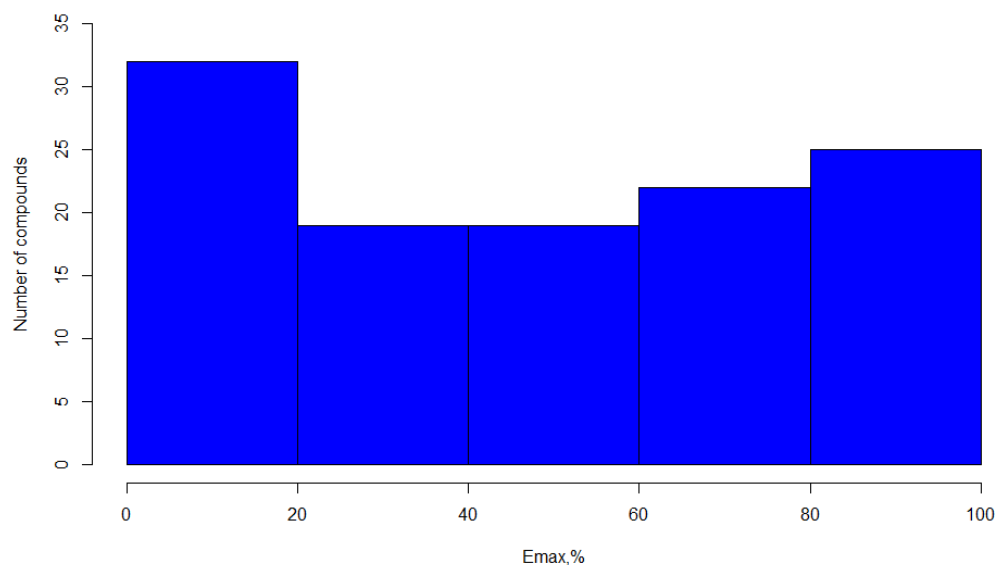


Figure 22. Data distribution for maximum antiviral effect (blue) and concentration which ensures it (red)

In order to build a QSAR model, E_{max} values were converted into class values, when 62 compounds with $E_{max} \geq 50\%$ were considered highly active and other 55 with $E_{max} < 50\%$ were considered inactive. ChemAxon Standardizer software [126] was used to apply rules for unambiguous representation of compounds structure, such as definition of major tautomer and the same way to represent functional groups. The protonation state

of every molecule major microspecies at pH=7.4 and all its stereoisomers were calculated for pharmacophore model developing using the ChemAxon *cxcalc* software [127].

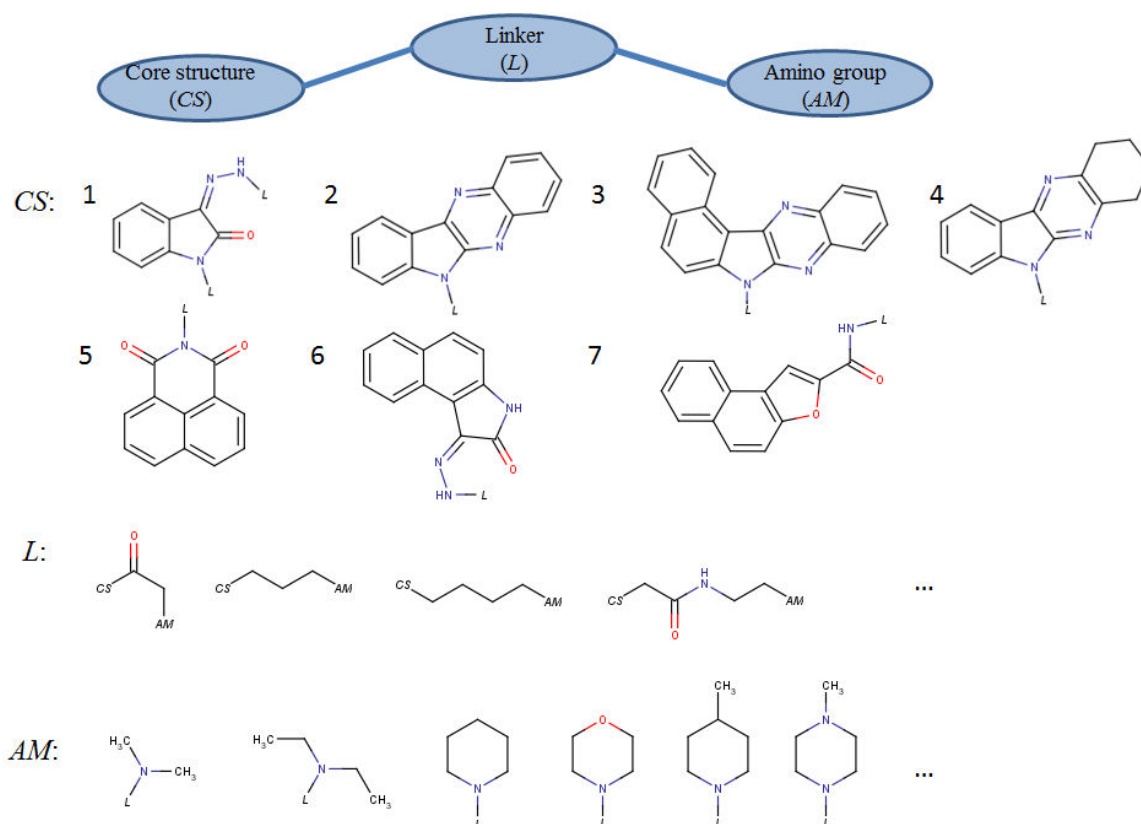


Figure 23. Description of the DNA binders used in the Training set

4.2 Modeling

The virtual screening funnel for antiviral compound screening consists of 3 tools: structure filters, pharmacophore models and QSAR models.

4.2.1 Filters

In chemoinformatics, a structure filter is a model in its simplest form, namely a rule (or ensemble of rules) based on selected structural features or physico-chemical parameters to assess whether a particular compound is eligible for further consideration. The number of fused rings (between 2 and 5), H-bond donors (0-3), H-bond acceptors (2-6) were used as structure filters due to the role of these parameters in intercalating activity. Parameters reflecting molecular flexibility (the number of rotatable bonds (3-12) and molecular weight (268-443) were also taken into account. The cutoff points

correspond to the minimal and maximal parameter values for the training set compounds obtained with the ChemAxon *cxcalc* plugin [127].

4.2.2 Pharmacophore models

LigandScout assigned pharmacophore labels and generated up to 500 conformers per every compound from 161 potent DNA intercalators. The compounds were clustered onto five groups containing from 5 to 99 compounds (**Table 4**) according to their pharmacophore patterns and conformers alignment score. Certain clusters contain representatives of more than one core structures, since the software found those compounds quite similar in terms of pharmacophore patterns. Then, approximately 30 – 60% of compounds, depending on cluster (totally 69 compounds) they were taken from, were randomly selected as the training set.

Table 4. DNA intercalators clusterization by LigandScout

cluster #	Number of compounds			Core Structure ^a
	training set	test set	overall	
1	4	1	5	#1
2	27	72	99	#2, #3
3	21	16	37	#5
4	3	2	5	#7
5	14	6	20	#1, #6
Total	69	97	166 ^b	

^[a] See Figure 23^[b] 5 compounds were chiral, therefore all stereoisomers were used as separate entities in model build

The resulting five pharmacophore models (one per each cluster) were validated on a test set composed of the remaining 97 DNA intercalators and 20000 decoys. Decoys were selected from the ZINC database according to the structure filters described in section 4.2.1. The only exception was the number of fused rings: since decoys must be structurally similar to the active compounds but void of key activity-defining features [102], decoy compounds were selected regardless of whether they comprise planar polycyclic system or not.

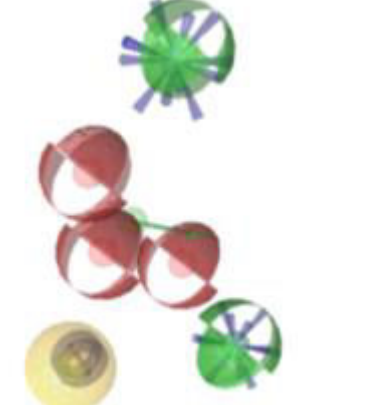
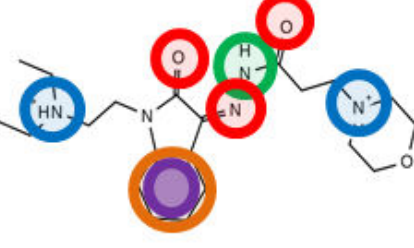



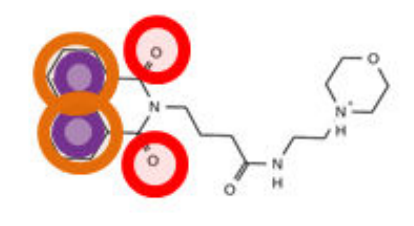
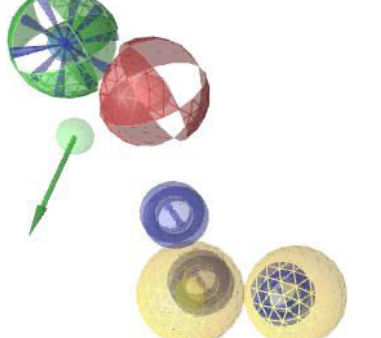


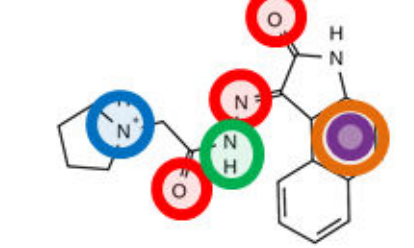
100% of actives (97 out of 97) were successfully retrieved during model validation. However, performance of models built on clusters 1 and 3 was dissatisfactory, since the number of decoys they retrieved is too high (Table 5). These 2 models were not used in virtual screening.

Table 5 Pharmacophore model validation summary

cluster #	number of compounds passed pharmacophore model		precision	recall
	actives	decoys		
1	1	16	0.06	1
2	72	0	1	1
3	16	134	0.11	1
4	2	1	0.66	1
5	6	0	1	1
total	97	151		

The failed models were built on clusters which consist of naphthalimide and disubstituted isatins derivatives (clusters 1 and 5 in Table 6, respectively). The models for cluster 2 (indolequinaxolines), 4 (naphthofurans) and 5 (monosubstituted isatins) were used in the virtual screening workflow.

Table 6 Pharmacophore models description

cluster #	pharmacophore model	example of compounds
1		
2		
3		
4		
5		

4.2.3 Classification SAR models

A 2-class (e. g. active and inactive class) model was built using antiviral data on hand of the 117 compounds described before. SiRMS and Random Forest were used as descriptors machine-learning method. Dataset was randomly spilt into the training and test sets, at an approximately 4:1 (94 compounds to 23 compounds) ratio, for model validation purposes.

Validation performance was acceptable (Table 7): Balanced accuracy for test set predictions is equal to 0.78 (Balanced accuracy = 1 for a perfect prediction) with neither active nor inactive compounds prediction accuracy falling below 0.5.

Table 7 Validation summary of classification antiviral model

	out-of-bag	test set
number of compounds	94	23
number of variables	60	
number of trees	250	
balanced accuracy	0.78	0.74
sensitivity	0.79	0.67
specificity	0.77	0.82
kappa	0.55	0.48

4.3 Virtual screening

Developed models were applied to screen a dataset of 3 207 605 compounds comprising BioinfoDB (3 207 317 compounds) and PCIdb (288 virtual compounds) according to workflow shown in Figure 24. At the first stage, the structure filters discarded the major part of compounds. Pharmacophore models were used for screening of the remaining 1 022 465 compounds, keeping 884 structures with DNA affinity potential (see examples in Figure 24).

It is worth mentioning that there were more than 884 compounds retrieved by pharmacophore models. Seven duplicates (e. g. compounds retrieved by several models at the same time) and seven rediscovered training set members were filtered out from the hits list. There was also an issue with some Model 5 hits related to imperfections in structure filters from Chemaxon. The software was only able to define fused ring substructure in the compound's structure yet intercalation requires planarity of polycyclic compounds' scaffold. Moreover, a potent DNA intercalation occurs when the planar

system contains at least 3 cycles. For example, in case of isatins the third cycle is formed by hydrogen bonding of hydrazone nitrogen and ketone group oxygen. Our pharmacophore models are unable to detect such niceties and define only aromatic and hydrophobic feature for isatin fused ring fragment, whereas structural filter minimum value for the number of the fused rings is set to 2. This leads to a possibility when compounds with a system of one planar and several non-planar fused cycles are retrieved by this model. Therefore, 87 hits retrieved by Model 5 had to be removed after visual inspection.

At the next step, remaining compounds were screened with two models: the antiviral SAR model reported in section 4.2.3 and the previously developed QSPR model for aqueous solubility prediction [84]. As a result, 87 potential antivirals (32 from BioinfoDB and 55 from in-house library) for which predicted solubility in water was larger than 10^{-5} mol/l were selected. All in-house library compounds passed applicability domain criteria, however 4 out of 32 compounds from BioinfoDB were out of AD. Additional predictions with the PASS software didn't display any adverse effects, toxicity and mutagenicity in discovered hits. Also, the search of the PubChem database revealed that none of the successfully screened compounds were previously used in antiviral bioassays.

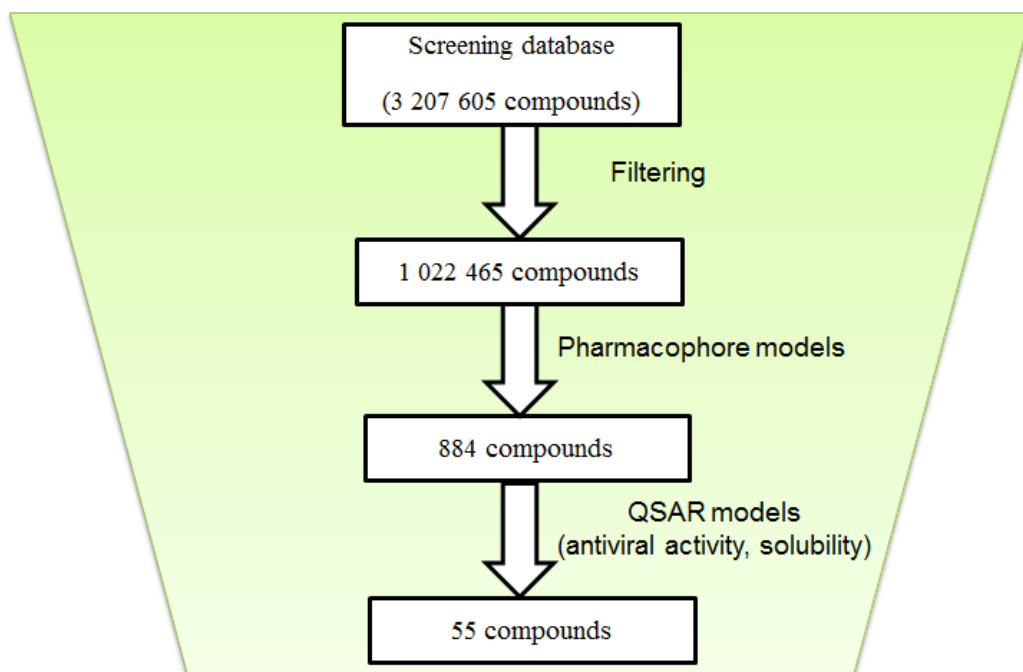


Figure 24. Virtual screening workflow.

Virtual screening of commercial databases, by contrast to on-purpose designed focused libraries may be more beneficial if the commercial compounds are actually available, i. e. they satisfy certain conditions, such as reasonable pricing, shipping delays, purity etc. If above conditions are fulfilled, using already synthesized compounds is more appealing than thinking about new synthetic methods for creation of entirely novel substances.

The other very important benefit from the virtual screening of commercial databases is indirect check if both the models and the additional in-house virtual libraries screened are good for usage. On the one hand, if a virtual screening procedure is applied to the commercial databases and it returns many more compounds than the typical hit rate for blind High Throughput Screening campaigns (consider an optimistic 0.5% as upper threshold), the virtual screening tools are clearly too permissive, and should be discarded – this was clearly not the case here. On the other hand, if hit compounds are found in the commercial databases, but the in-house designed library does not yield any, it means that the design of the latter failed to focus on relevant structural features which must not be accepted for such database. Therefore, combining the commercial and in-house sources of compounds in virtual screening is important, even if it requires additional computational. Approximately 20% of in-house database and 0.001% of BioinfoDB compounds were considered to be virtual hits. This illustrates the both the efficiency of the virtual screening protocol and proper design of the in-house library with the focus on molecular fragments known to ensure antiviral activity compared to non-specific drug-like compounds library.

Thus, in light of above-discussed pros and cons, priority was given to the synthesis and biological testing of 55 in-house hits.

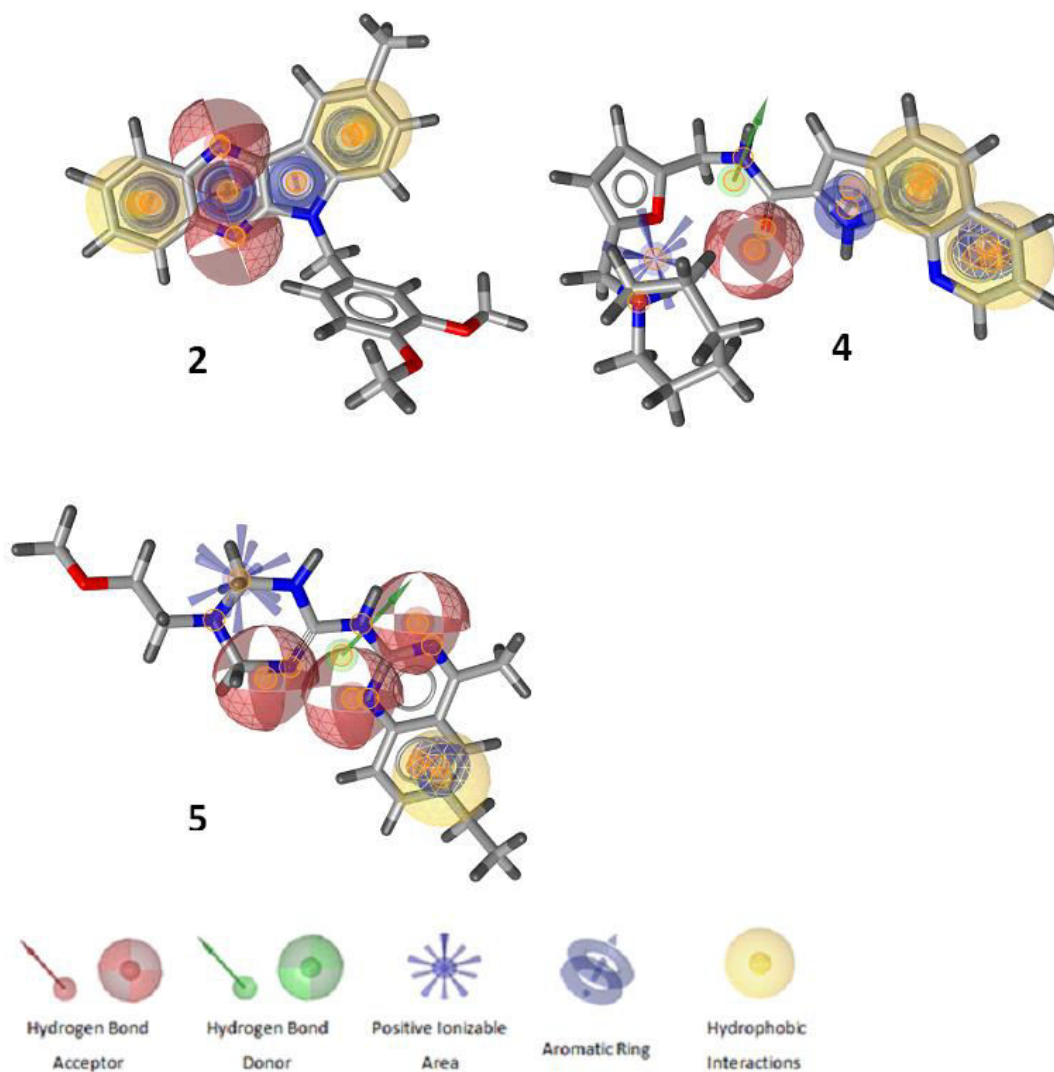


Figure 25. Example of compounds selected in virtual screening aligned with pharmacophore models. The number near the structure corresponds to the cluster on which the model has been developed (Table 4).

Synthesis of virtual hits and antiviral activity measures were carried out at A.V. Bogatsky Physico-Chemical Institute (PCI) NAS of Ukraine in Odessa and the Institute of Chemical Biology and Fundamental Medicine (ICBFM) in Novosibirsk, Russia, respectively. Detailed description of this can be found in Appendix B Table S5, Figure S1. Inhibition of *Vaccinia virus* reproduction was carried out in CV-1 cells. 40 out of 55 compounds were soluble enough to be used in GFP-based antiviral bioassay. Among these compounds only 5 have shown efficiency in GFP inhibition and, therefore, were chosen for the viral plaque assays. As a result of the latter assay, 2 compounds displayed

high activity at concentration below acute toxicity levels (Table 8). The most active compounds were tested in interferon induction assays and showed no induction capacity.

Table 8 Antiviral activity of potent compounds measured by classical plaque forming assay.

Compound ID	C, μM	Viable cells, %	Virus titer, lg (PFU/ml)
K⁺ ¹⁾		100 \pm 0.1	3.3 \pm 0.1
10	1	107.8 \pm 0.1	3.1 \pm 0.1
	10	96.9 \pm 0.2	2.8 \pm 0.1
	50	75.6 \pm 0.1	2.4 \pm 0.1
24	1	103.6 \pm 0.2	3.4 \pm 0.1
	10	131.4 \pm 0.1	2.8 \pm 0.1
	50	90.0 \pm 0.1	2.5 \pm 0.0

n.d. – activity not determined. 1) Virus titer in the infected cell incubated in the presence of 0.002, 0.02 or 0.1% of DMSO in the cell medium (correspond to the DMSO concentration in the medium with 1, 10 or 50 μM of the compounds) was 3.1 \pm 0.3 PFU/ml, similar to K⁺.

4.4 Conclusions

Two most active compounds (Figure 26) **10** and **24** inhibit virus reproduction by at least 8 and 6 folds, respectively in considerably lower concentrations than their CC_{50} , which makes them eligible candidate for further antiviral research. The discovered hits were tested for DNA affinity according to the procedure reported in [125]. They display reasonable intercalating activity: $\lg(K_i) = 6.03$ and 5.20 for compounds **10** and **24**, respectively. Considering the absence of interferon induction, the results are consistent with the basic hypothesis of this study that broad-spectrum antiviral activity is linked to nucleic acid intercalation.

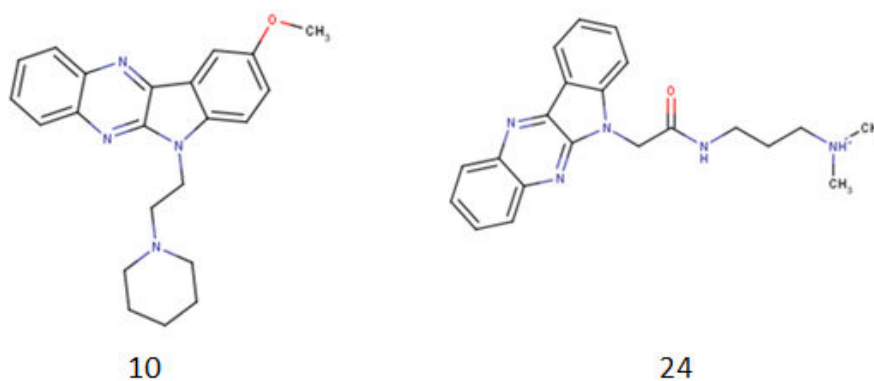


Figure 26. Prospective antiviral agent candidates

PART 5 VISUALIZATION AND ANALYSIS OF CHEMICAL SPACE OF ANTIVIRAL COMPOUNDS

In computer-aided drug design of new nucleic acid intercalators reported in Part 4, two compounds with high activity against Vaccinia virus and acceptable toxicity were discovered. Determining whether they possess activity against any other virus requires some additional experimental investigations which is out of the scope of this work. On the other hand, some suggestions about potential pathogen targets could be made using the data on known antiviral compounds. In this part of thesis, we report chemical space visualization and analysis based on the GTM approach.

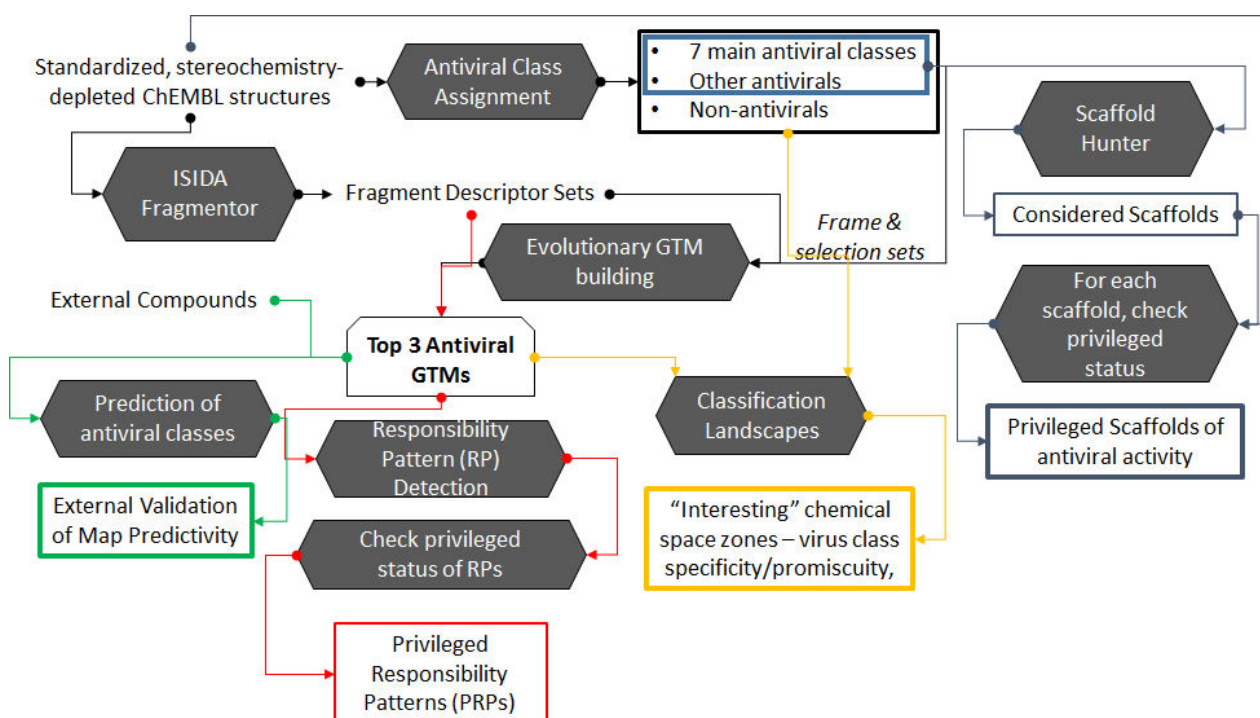


Figure 27 Workflow of Chemical space analysis and visualization

This work has several objectives. The first one is *in silico* prediction of an antiviral activity profile – including antiviral propensities for as many viruses as possible as profile components. Such prognosis can be reliable only if enough data on antiviral activity is collected and an appropriate algorithm is used for model building. The second one is finding areas of the chemical saturated with antiviral compounds active against particular type of virus.

The workflow (Figure 27) consists of four major parts:

- Data curation
- Model development

- Chemical space analysis
- Activity prediction of external compounds

5.1 Data curation

An alleged analysis of the entire chemical space of relevance for antiviral design must be founded on rich, accurate and diverse structure-activity information. Even though commercial antiviral compounds databases exist [128], there is still considerable free access structure-activity data. Systematization and thoughtful description of this data was crucial to the further analysis. Firstly, an extensive data extraction and curation from the heterogeneous, multi-source activity data was followed by compound structure cleaning, standardization and duplicate removal. The ChEMBL database [107] was chosen as a reliable source of publicly available bioactive compounds for the current study. The query result was downloaded as a CSV file containing 52 columns with data description (e. g. Assay Organism, Compound's ID, Compounds' Canonical Smiles etc.) and 114 324 rows of data entries, corresponding to 35 547 compounds which have distinct ChEMBL IDs. Activity data extraction and curation was performed with the KNIME [108] software. It was used to implement filtering rules in data table from above-mentioned CSV file. The rules for removing entries in case of inconsistent data were as follows:

- **Data validity comment** contains one of these keywords: "Outside typical range", "Potential missing data", "Non-standard unit for type". For example, in one of the original articles [129] compound "11e" was assigned activity value $> \sim 0.3 \mu\text{M}$. The author decided that this compound is inactive. Therefore, above-mentioned type of data was excluded and 33 370 compounds (distinct ChEMBL IDs) remained.
- **Activity comment** does not report "active", further decreasing compounds number to 32 431
- **Potential Duplicate** column signals redundant data. For instance, in [130] some compounds oddly have the identical activity value against 3 different strains of HIV. This can be the sign of data being simply copied and pasted to ChEMBL. Therefore, data labelled "Duplicates" were removed. – 32 420 compounds remained.
- **Assay type** contains "ADME" (thus reporting pharmacokinetics data on the host, by contrast to functional [131] and binding [132] assays, which were kept – 32 373 molecules).
- **Assay CRC description** was not set to "Scientific Literature". In this study, we emphasize the importance of keeping results published in scientific literature over the ones published in sometimes classified bioassays (e. g. PubChem assays) because

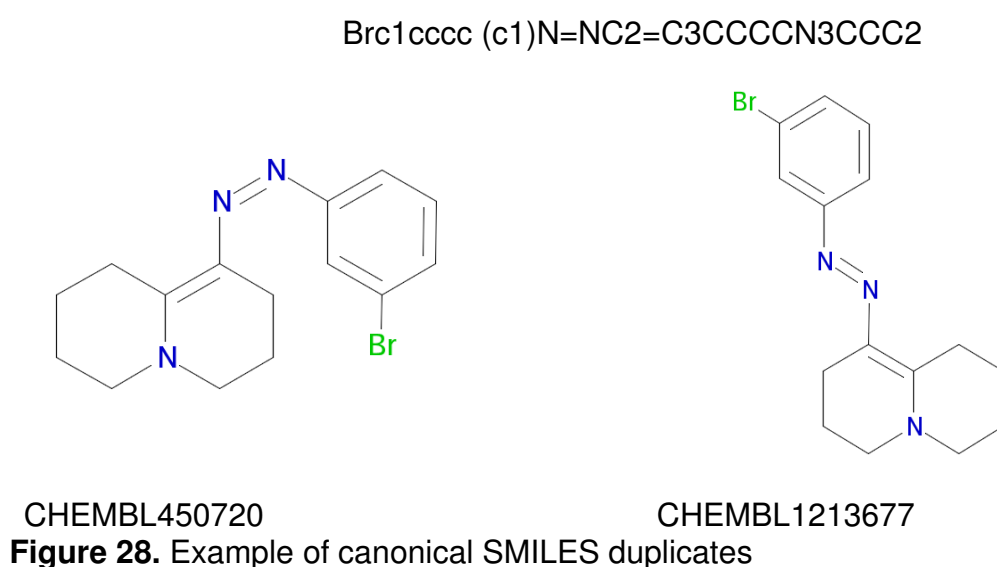
data published in the article (book, thesis etc.) can be traced back from the database entry to the original source revealing all the details about bioassays, compounds etc. – 32 348 compounds remained.

- **Standard type**, activity values can be a result of rare type activity measurements such as replication efficiency, [133, 134] or activity parameters that were not actually related to antiviral activity. More detailed information concerning this field is given in Table 9, together with the specific thresholds defining “active” status with respect to the **Standard value** field, which must be interpreted in relation to the **Standard type** and **Standard units** fields. A total of 24 629 molecules were still selected after this last stage.

Table 9 A curated database was created using the following criteria. Standard type, Relations, Standard value and Standard units are column names in ChEMBL database file.

Standard type	Relations & Standard value	Standard units
Activity	≤ 100000	nM
EC50	≤ 100000	nM
EC50	≤ 50	µg.mL-1
ED50	≤ 15	mg ml-1
ED50	≤ 100	mg.kg-1
ED50	≤ 100000	nM
ED50	≤ 1000	µg ml-1
ED50	≤ 100	µM
ED50	≤ 100	µmol.kg-1
Emax	> 50	%
IC50	≤ 100000	nM
IC50	≤ 50	µg.mL-1
IC90	≤ 100000	nM
IC90	≤ 50	µg.mL-1
Inhibition	> 50	%

Activity data curation via KNIME yielded 49 191 reliable data entries, comprising information on 24 629 unique ChEMBL SMILES strings. This corresponds to 24 633 different compounds (in the sense of distinct ChEMBL compound ID values), published in 1982 articles. Four compounds [135,136] were erroneously duplicated in ChEMBL, and given different ChEMBL ID for the same structures. In fact, researchers from [136] have simply used [135] data for computational studies and have clearly referred to the data source in their paper. Apparently, they mistook different two-dimensional structure representation for double-bond stereoisomerism (Figure 28), even though authors never mentioned anything about compounds stereoisomers.



After activity curation was finished, the structure standardizing using the in-house rules implemented on our virtual screening web server was carried out. This procedure was powered by the ChemAxon [126] toolkit and included the following steps:

- removal of compounds containing heavy metal species and >100 heavy atoms
- salt removal,
- inorganic compounds removal
- conversion into the (predicted) most stable tautomer form,
- representation of N oxides with split formal charges, conversion to the “basic” aromatic forms of 5 and 6-membered aromatic rings, etc.) ...

Since we decided to use a 2D approach for chemical space description, some of the unique SMILES strings may be linked to several ChEMBL ID values – which may correspond to different stereoisomers mapped onto a common stereochemistry-depleted

SMILES, to different formulations (counter-ions in salts) accompanying the same active principle, or simply to genuine duplicates of a same structure under different ChEMBL IDs in the original database.

The next step was definition of the major antiviral classes. It required an analysis of all the different antiviral assay protocols reported in the filtered entries, in order to define a clear grouping criterion to associate activities from particular assays with antiviral classes. In such way, each compound found active in a particular assay becomes a “positive” representative of the class to which the assay was assigned. Since the viral protein targets variety is enormous, but many antivirals lack a definite mechanism of action, virus type was chosen to be the criterion for grouping antiviral compounds.

The International Committee on Taxonomy of Virus (ICTV) developed a virus classification which we used in this study [137]. The most important thing was to choose the appropriate classification taxon among existing taxonomic ranks from **Order** to **Species**, since **Strains** are too specific to design the antiviral compounds against. It is well to bear in mind that viral taxonomy is a very complicated thing because viruses are relatively simple organisms and even slight changes in their structure can lead to big change in functionality, therefore, it is hard to define the clear difference between them. The most unambiguous taxonomic rank for virus with maximum discrimination is **Family** since it is based on criteria, such as:

- Genome nature (e. g. dsDNA, ssRNA (-))
- Envelope (presence or absence)
- Morphology (i. e. virion form)
- Virion form (e. g. bullet-shape)
- Genome configuration
- Genome size
- Type of host organism

However, **Family** is insufficient for grouping since none of the mentioned criteria takes into account protein composition of the virus, making structure-activity determination more complicated. Therefore, lower levels of hierarchy, namely **Genus** and **Species** were examined. Further consideration revealed that both of these taxonomic ranks take virion

protein composition into account but **Genus** allows grouping more viruses of similar origin into one category (e. g. HSV-1 and HSV-2 have similar protein composition but they are assigned to different **Species** since HSV-1 causes mostly sore colds and HSV-2 causes mostly genital herpes) [138]. Thus, virus **Genus** was chosen as a classification criterion, resulting in the definition of 7 major activity classes. **Genera** of major classes include only mammalian viruses.

Data from ChEMBL has the Information on virus types in Assay Organism column. Entries with the Assay Organism field matching one of the seven text-mining queries below (an asterisk matching any pre- or postfix characters) were assigned into corresponding antiviral classes (details given in Table 10 below).

Table 10 Text queries used to classify ChEMBL compound-activity records into antiviral classes, on hand of the Assay Organism entry. The “Total antivirals” number is lower than the sum of listed class members, because some compounds may be members of several classes.

Antiviral Class	Assay Organism matches:	Hit count
Enterovirus (<i>Ent</i>)	“*Human rhinovirus*” OR “*Human enterovirus*”	424
Hepacivirus (<i>Hep</i>)	“*Hepatitis C virus*”	5320
Influenza A (<i>Inf</i>)	“*H2N2 subtype*” OR “*Influenza A virus*”	638
Lentivirus (<i>Len</i>)	“*HIV*” OR “*Human immunodeficiency virus*”	8854
Orthohepadnavirus (<i>Ort</i>)	“*HBV genotype D*” OR “*Hepatitis B virus*”	700
Pestivirus (<i>Pes</i>)	“*Bovine viral diarrhea virus 1*”	412
Simplexvirus (<i>Sim</i>)	“*Human herpesvirus 1*” OR “*Human herpesvirus 2*” OR “*Hsv-2*” OR “Herpes simplex virus (type 1 / strain F)*”	790
Other antivirals	Entries not matching any of the above	7897
Total antivirals	Sum of above, compounds present in several classes counted only once	24629

This approach allows linking ChEMBL compound IDs to the seven specific antiviral classes, plus the “other antiviral” class containing antiviral agents against any non-mentioned virus **Genera**. If, a specific ChEMBL ID was associated with one (or more) of the seven major viral classes, it was labelled as “positive” with respect to the class(es). The “negative” status with respect to a class was assigned to a given ChEMBL ID if it represents a positive associated with another class or it has not been recognized as

positive at all. Compounds labelled “other antivirals” systematically appear amongst the “negatives” associated with the seven main classes.

The data extracted from ChEMBL is very heterogenous, since experiments described in 1982 distinct papers were carried out by different research teams using different bioassay protocols. Therefore, grouping of antivirals was a challenge. For this reason, but also because the latter strategy leads to larger data sets (providing much-needed statistical robustness for further analysis), a higher-level merging of ChEMBL sets – by viral class membership and irrespective of specific assay conditions – has been preferred in this work.

As it was mentioned before, all antiviral compounds were considered active against some particular virus and inactive against others. However, in order to thoroughly analyze the key structural features of antivirals and create a valid tool for activity prediction using mapping techniques, comparing them against completely inactive compounds was of great importance. For this reason, compounds with no record of antiviral activity from ChEMBL database with all structure standardization procedure applied to them were added to actives making up to 1.2 million substances.

Each unique standardized and stereochemistry-depleted compound structure (SMILES string) [139] ended up corresponding to the (one or several) ChEMBL IDs. For each of the 1.2M standardized compounds, the ChEMBL IDs were searched within the listed “positives” and “negatives” associated with each of the seven virus classes. If none of the ChEMBL IDs of a standard compound is present in that list, that compound was classified as “outside” the antiviral chemical space, and labeled “0”. If at least one of the ChEMBL IDs was present amongst the entries of one out of 7 classes, then the compound was labeled as positive with respect to that class. Negatives of each class are, by contrast, all the positives of other classes – except for the “promiscuous” compounds which were active against several major classes of viruses– and the “other antivirals”. Eventually, an antiviral profile text file has been compiled for the entire 1.2M ChEMBL collection of standardized, stereochemistry-depleted SMILES strings. It is a seven-column file, each line corresponding to a structure M and Each column corresponds to an antiviral class C, in alphabetical order as given in Table 10. Status labels in this matrix, Stat (M,C) may be “2” if M is a positive of class C, “1” if M is a negative of C, or “0” if M is a structure outside of the antiviral chemical space – in this case, Stat (M,C)=0 \forall C.

It is important to note upfront that a categorization as “positive” is a clear statement that this substance has an antiviral effect on at least one member of the given viral class. “Negative” in most cases mean “unknown activity” for that class, since most compounds were never tested against huge variety of viruses. The “negatives” are to be used as examples of compounds associated with different viral classes, in an attempt to learn what features differentiate the drug candidates of one class of viruses from those associated with another. While “positive” is synonymous to “active”, it is not reasonable to interpret “negative” as “inactive”, especially in this context where labels do not refer to a specific viral strain, but to a whole class. Technically, a compound could be declared “inactive” against a group only if it would be tested and found inactive against each virus of that group – an impossible endeavor. Also, note that ChEMBL compounds which were never reported to participate in any antiviral tests, and herein considered “outside” of antiviral chemical space are distinguished from “negatives”, even though they are also likely to be inactive. The difference is that a true “negative” has the peculiarity to be considered, by at least one group of scientists, as relevant enough in order to deserve being screened against at least one viral strain. In our case “Negative” should be interpreted as “presumably different” from the effective antivirals of given virus class, all while being interesting antiviral compounds, targeting other viruses. Therefore, if, for example, a standardized, stereochemistry-depleted structure is associated with two ChEBML IDS, one of which (ID1) is reported as “positive” against viral group 1, while ID2 is given as “positive” against another group 2, careful investigation is needed. Apparently, this is a paradoxical situation, since ID1 as “positive” for group 1, and not encountered in any measures run against group 2, would be by definition labeled as a “negative” of group 2, and vice versa. Or, both IDs refer to a common standardized structure, i.e. to a common point in the vector space of stereochemistry-ignorant molecular descriptors. If ID1 and ID2 represent a case of genuine compound deduplication (compound has no stereoisomers, but perhaps comes in different formulations, as salts with different counterions, etc.), it is safe to assume that, whilst some authors have used the substance ID1 to report their test on class 1, the test on class 2 running under ID2 concerned exactly the same compound. Therefore, it is safe to decide that the compound is “positive” with respect to both classes, the apparent problem being due to ChEMBL, having referred to it by different IDs. If, however, the compound has stereoisomers, one should check if the two different testing protocols did actually involve the same isomer. However, the goal of this research is not to check ChEMBL database quality, nor to investigate aspects of stereochemistry, but to define the robust structural features associated with antiviral

activity. After structure standardization and generation of stereochemistry-depleted unique SMILES codes – only 2.5% of compounds were associated with more than one compound ChEMBL IDs, i.e. represent potential stereochemistry-related issues. This number is conveniently small to justify the usage of 2D molecular descriptors in this work.

The herein employed classification scheme – “positives” vs. “negatives” for each virus class, plus ChEMBL compounds “outside” antiviral chemical space is not an experimental activity profile matrix, in which each compound×target table cell represents a measured activity value. Such a matrix should have had specific virus strains listed as targets, and would have been extremely sparse, thus virtually useless for robust analysis of structure-activity trends. The classification advocated here is not a compounds activity profile, but a snapshot, prone to further evolution, of what is known for sure to work and what medicinal chemists would expect to work against virus groups. This coarse view has the merit of robustness, and if chemoinformatics may prove that the above-mentioned chemical subspaces are well distinct, it could allow to outline the path for further antiviral drug design.

5.2 Model development

The machine-learning algorithm used in this study is Generative Topographic Mapping (GTM) [71,72,140,141]. Optimally discriminating GTMs were built following the same evolutionary strategy [142] used to generate “universal” maps of maximal generality for the entire drug space.

Descriptor selection is a very important part of building both descriptive and predictive GTM. previously used in GTM-related studies [142] 38 different ISIDA fragmentation schemes provided descriptor choices for the evolutionary algorithm, i. e. Darwinian selection procedure. Evolutionary algorithm encouraged the selection of the fragmentation schemes which were able to successfully discriminate compounds of 7 selection sets featured the actives of each class, by contrast to other class members and “other antivirals”.

Frame sets are compound collections used to generate the GTM manifold. Thus, they need to be chosen such as to span the entire relevant zone of the chemical space, therefore providing points of support for a robust fitting of the manifold. Here, subsets of the antiviral set, of various sizes, were taken, independent of compound assignment to antiviral classes (frame sets do not convey any activity-related information, since manifold construction is completely unsupervised). As automatic frame set selection is also a

degree of freedom of the evolutionary map building procedure, three different frame set choices were considered in this case: (1) the entire antiviral set, (2) half of the antiviral set (every second entry) and (3) a quarter of the viral set (one compound out of four).

During the evolutionary procedure, the fitness score used for map selection was based on the 3-fold cross-validated capacity to discriminate positives of each of the seven antiviral classes from its negatives (which make up the rest of the antiviral set). In other words, the seven different “selection sets” were represented by the same antiviral compound set, but in association to the seven different status labels. As the selection is driven by success in seven different classification tasks, the overall success score (map fitness score) is calculated on the basis of individual balanced accuracies for each task, as their mean value penalized by their standard deviations.

Out of the top maps emerging from the Darwinian evolution simulation, the best three based on distinct descriptor choices were selected. Integer responsibility patterns for each antiviral compound were determined on each map according to equation (19). Likewise, ChEMBL molecules outside the antiviral space (labelled class “0”) were retrospectively mapped, and their responsibility patterns extracted.

Three top performing antiviral maps – in terms of simultaneously discriminating positives from negatives, for all the seven antiviral classes, in a 3-fold cross-validated prediction run – are depicted in Table 11. They were chosen to represent different views on chemical space, have various sizes, but are all successful in solving the above-mentioned cross-validated discrimination problem, meaning that there is no unique recipe to capture the chemical information associated with antiviral activities.

Table 11 GTM parameters. ISIDA fragmentation schemes are (a) circular atom&bond fragments of size (topological radii) between 1 and 3, (b) circular pair counts of sizes 1 to 5 and (c) circular atom &bond fragments colored by the CVFF force field type. Map sizes (n) are reported as numbers of nodes per line of the square grid. m - number of RBF centers, w - RBF width factor, l - regularization coefficient.

Map	Descriptors	Size	m	w	l
1	IIAB-1-3 ^a	28x28	18	1.5	8.128305
2	IIRA-P-1-5 ^b	34x34	25	0.4	0.087096
3	IIAB-FF-1-2 ^c	42x42	30	0.9	19.952623

All the maps manage to highlight specific chemical space zones associated with each of the considered classes, with no balanced accuracy scoring below a respectable

0.77. Distinction between the positives and negatives is most difficult for the lentivirus class, which is also the richest one in terms of associated positives.

Table 12 comprises the information on averaged statistical parameters of 3-fold cross-validation for every model, including the fraction of correctly classified positives (“sensitivity”) and negatives (“specificity”) that compose the BA score.

Table 12 Top antiviral maps emerged from Darwinian optimization. Winning ISIDA fragmentation schemes are (a) circular atom&bond fragments of size (topological radii) between 1 and 3, (b) circular pair counts of sizes 1 to 5 and (c) circular atom &bond fragments coloured by the CVFF force field type. For more detail, see ISIDA fragmentation scheme nomenclature. [143]

Map	Descriptors	Size	Cross-validated Balanced Accuracies/antiviral class						
			<i>Ent</i>	<i>Hep</i>	<i>Inf</i>	<i>Len</i>	<i>Ort</i>	<i>Pes</i>	<i>Sim</i>
1	IIAB-1-3 ^a	28x28	0.90	0.83	0.82	0.79	0.83	0.86	0.82
2	IIRA-P-1-5 ^b	34x34	0.89	0.84	0.81	0.77	0.83	0.85	0.80
3	IIAB-FF-1-2 ^c	42x42	0.91	0.83	0.79	0.77	0.80	0.86	0.83

Validation of the GTM building process is conceptually more complex than the one of a typical regression or classification model, because it includes several distinct steps. First, manifold construction is totally unsupervised, already published results [142] show that being part of the frame set serving for manifold fitting is not enhancing the quality of prediction of such compounds. Next, a given manifold needs to be “coloured” by a property, using a training set of compounds, and the resulting property or class landscape may serve for prediction of external compounds. By default, training and external compounds may be obtained by splitting the complete pool of available structure-property information into (typically) 2/3 for training and 1/3 for external prediction. Iteratively, each tier plays the role of test set, being subject of antiviral (positive/negative) status prediction, by projection on the map coloured by the other two tiers. This test set is external, because it never contributed to colour the underlying map. The dataset is shuffled and the procedure is repeated three times. Reshuffling and repeating ensures that the prediction outcomes are not biased by any peculiarly favourable regrouping of compounds in test and training tiers. This “aggressive” triplicated 3-fold cross-validation (XV) adopted in this work is simply the more rigorous alternative to classical external testing on a single test set due to the lack of free-of-charge medicinal chemistry data not covered by ChEMBL.

In terms of computational effort, triplicated 3-fold XV amounts thus to nine GTM “colouring”/prediction cycles, so takes roughly the same times as a nine-fold classical XV, all while being both a much more challenging exercise – because it minimizes the information effectively used for model learning and thus maximizes the opportunities for misprediction.

Triplicated 3-fold XV is the source of map goodness (“fitness”, in the evolutionary context). Therefore, the entire available antiviral SAR information extracted from ChEMBL was used for map selection. The selected maps above are the maps that maximize predictive power, in terms of separation propensities of the considered antiviral classes, in the context of aggressive, triplicate 3-fold cross-validation. It is thus justified to ask the question: is there a risk of “overfitting” by throwing the entire SAR information into the map selection process? In this context, “overfitting” means that the maps perform well on the current SAR data only because they were selected to perform well with respect to them. Allegedly, they may not perform as well on different antiviral compound collections. However, we do not dispose of an independent SAR data set of comparable size and richness to directly challenge this issue.

3-fold XV does assess model robustness but a better way to look at its predictive capacity is to predict activity for an external test set. A GTM colored by node-specific predominance of positives vs negatives of a given antiviral class may be used as a predictor of the category to which a novel, so far unreported compound is most likely to belong. This is achieved by positioning the novel compound on the colored map, and reading out the locally predominant class at its residence point. The three antiviral maps built in this work were challenged to predict compounds of known antiviral class association, but not accounted for at the training stage. Their antiviral data was published in *Antiviral Research*, a journal which seemingly was not in the scope of ChEMBL’s data mining. Compounds are active against HIV [144] HCV [145,146] HSV [147] and Influenza A [148,149,150,151] virus. Test set compounds are considered to be predicted positives of a particular class if at least two of the three maps position them in positive-dominated class landscape zones.

All 3 models were validated by external test set. Unfortunately, the quest for genuinely “external” validation data in the above sense was of rather limited success. The additional data used to challenge the maps in an external antiviral class prediction exercise consisted of 10 anti-Influenza virus A, 2 anti-lentivirus, 5 anti-hepacivirus and 2

anti-simplexvirus compounds. Except for the Influenza A subset, these numbers are too scarce for robust validation (a state of fact showing how difficult it may be, in practice, to find experimental data not yet part of ChEMBL), but prediction was attempted nevertheless. Results were excellent for the Influenza A subset, where 9 out of 10 candidates were correctly recognized as positives. Also, both anti-lentivirus compounds have been recognized, whilst, however, none of anti-simplex or anti-hepacivirus candidates were predicted positive.

5.3 Chemical space analysis

5.3.1 Visualization of the chemical space

In the present work, various classification landscapes were generated. First, antiviral class-related landscapes distinguishing between positives (“2”) and negatives (“1”) for each of the seven antiviral classes will also be used for predicting the estimated class membership of novel compounds, thus providing a mechanism for external model validation. However, formal class labels 1 and 2 can be reassigned in order to monitor the generic chemical space occupied by the entire antiviral set (now collectively assigned to class “2”) by contrast to the rest of the ChEMBL database (outside of antiviral chemical space, now class “1”). Single class plots may also be realized, when the only variable is density.

Figure 29 represents the classification landscape of the lentivirus-positives by contrast to the rest of the antiviral compounds, and it clearly displays the multiple blue zones in which lentivirus-positives “cluster” together on the map. The existence of such zones is a consequence of the high balanced accuracy values (a map with no discriminating power would be entirely coloured in yellow-green, and return a balanced accuracy score about 0.5). However, these multiple zones are scattered all over the relevant chemical space, signalling that lentivirus-positives for a large and very diverse collection of different compounds, targeting different antiviral mechanisms.

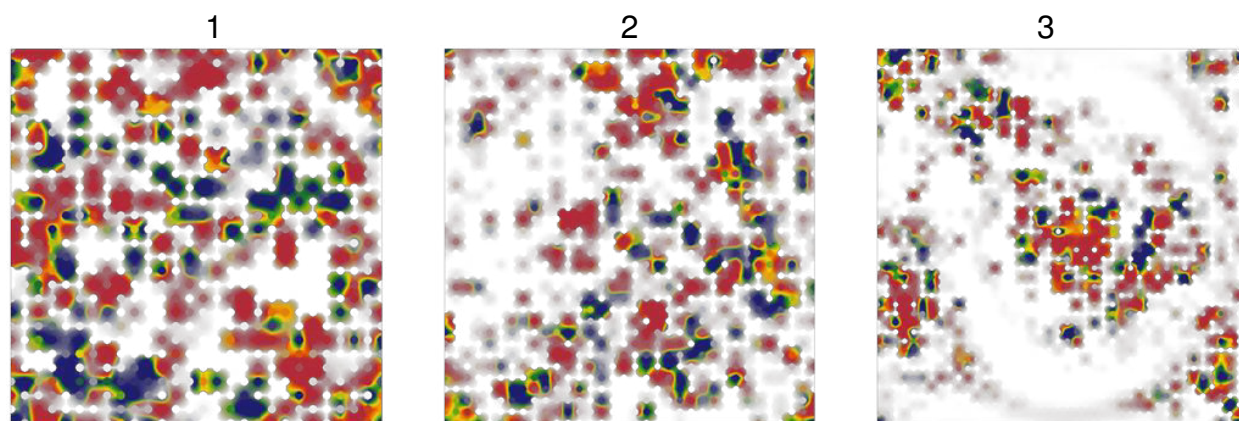


Figure 29. Classification landscape of Lentivirus-positives (class “2”, blue) vs Lentivirus-negatives (class “1”, red) on the three antiviral maps. The intermediate colours are as follows: red 0-0.2, orange 0.2-0.4, yellow 0.4-0.6, green 0.6-0.8, blue 0.8-1.0 where number is probability of finding an active compound residing within the node. Maps are numbered according to Table 12.

Unfortunately, the ChEMBL database does not report a sufficiently large series of viral protein inhibition tests, which might have helped assigning the various “lentivirus islands” on the map to different mechanisms of action (if viral target inhibitors would be found to reside within above observed islands). Failure to retrieve sufficient in vitro activity data against virus targets (including well-known proteins such as HIV protease and reverse transcriptase) from ChEMBL shows that the present-day antiviral compound research has been driven forward mostly by anti-pathogen tests. Specific assays aimed at understanding the interactions with target proteins were realized only for few validated leads or short series of analogues, in Medicinal chemistry sense of this term. However, such compound sets are small, biased, and do not support global statements with respect to the entire antiviral chemical space.

The other interesting observation from Figure 29 is that the increase of map resolution (grid size) translates to an increase of the number of marginally populated nodes, and not to a better distribution of the antivirals over more nodes. This is not surprising, since the increase of the map size was not followed by an increase in discriminating power, suggesting that the smaller map 1 is already sufficiently large to accommodate the chemical diversity spanned by the antiviral compounds.

Figure 30 is a comparative display of the classification landscapes for six out of seven (excluding the least numerous *Enterovirus*) antiviral classes of compounds on the map 3. All these represent the same global compound set – the entire antiviral set – in

which the space zones dominated by each of the six classes are, alternatively, highlighted. Positives of every antiviral class form chemically diverse collections, but each has a rather distinct scattering pattern on the map. Compounds associated with different virus classes show clear and distinct pictorial “signatures” on the map.

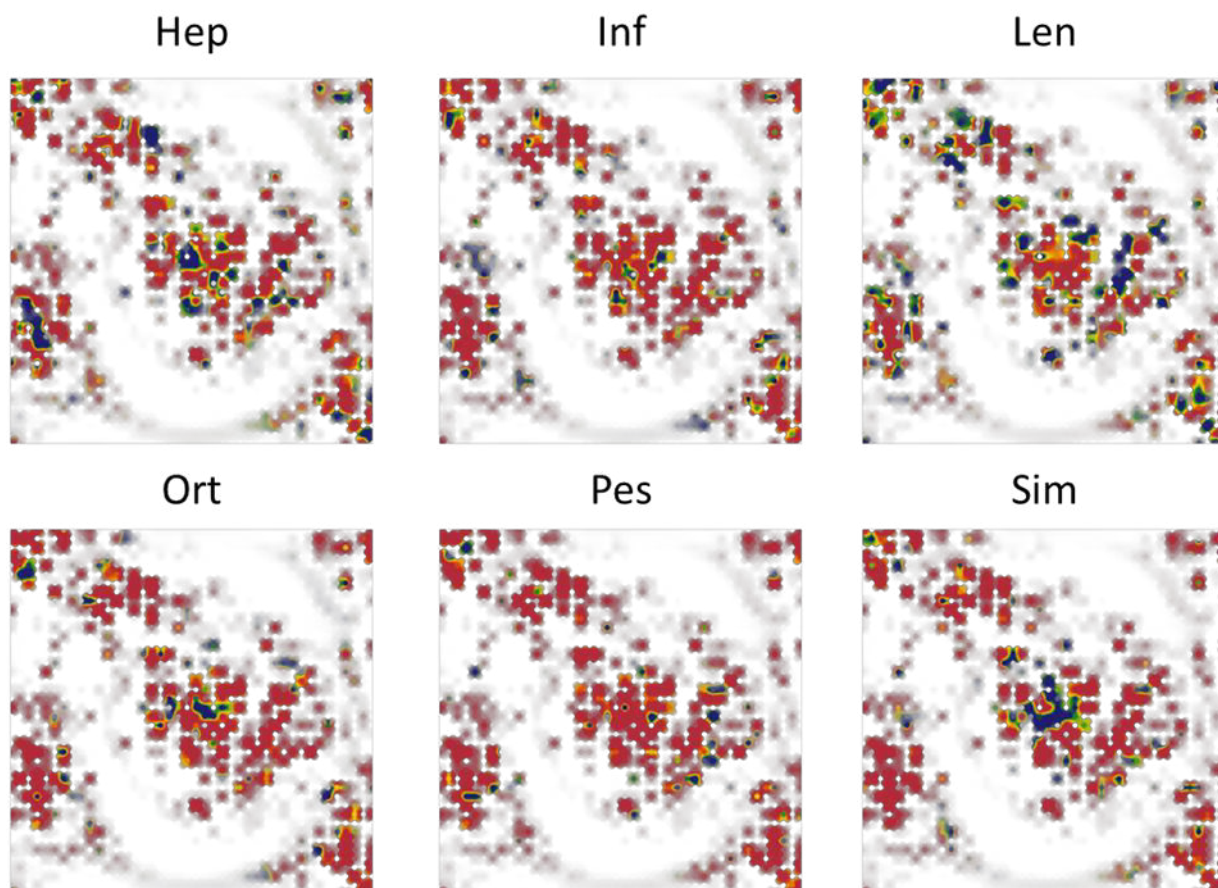


Figure 30. Classification landscapes for six of the seven virus classes, on map 3 (enterovirus, left out, is the smallest group mapping on few distinct nodes)

When classification landscapes like in Figure 29 are matched against one-class plots (positives of a given virus class) shown in Figure 30, it is possible to evidence the chemical space areas where the positives are outnumbered in terms of normalized density. The left-hand plot in Figure 31 represents the density trace of the Hep positives. If these would exclusively occupy chemical space zones void of, or sparsely populated by any other antivirals, then all the high-density areas on the left should match blue, Hep-dominated areas in the classification landscape right. This is mostly true – otherwise, no high balanced accuracy score could have been reached – but not always. Visually, it is

easy to pinpoint the areas in which significant subsets of Hep-positives are outnumbered by other antiviral compounds (three such spots were highlighted).

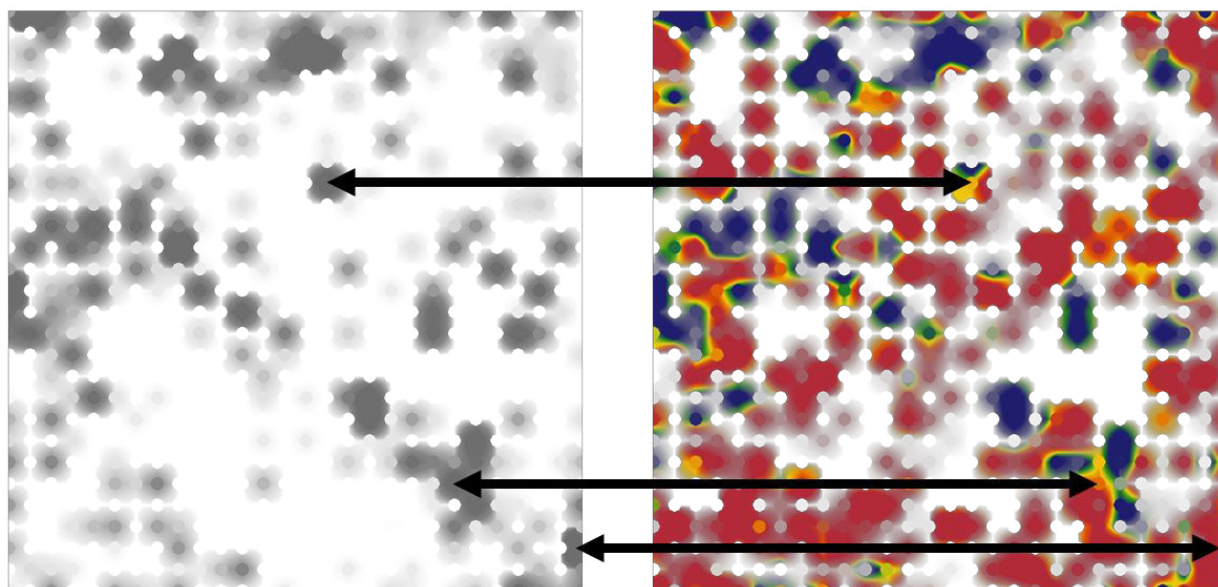
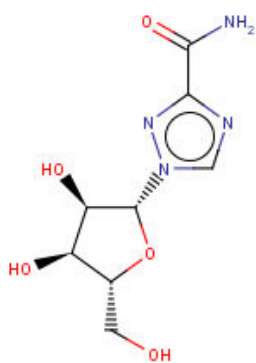
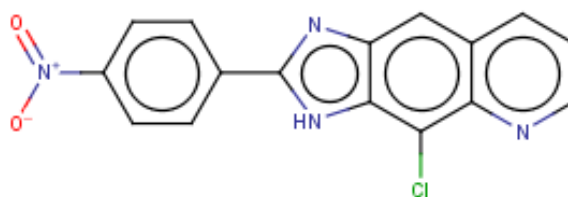


Figure 31. Density trace of Hep-positives (left) versus classification landscape of Hep-positives, as rendered by map 1. Arrows highlight areas that are densely populated by Hep-positives, but are dominated by antivirals of different classes.

There may be two alternative explanations for the existence of such mismatches: the pessimistic one is that the limit of accuracy of the GTM model is attained, while the optimistic one would be the claim that therein found Hep-negatives are actually not yet discovered actives. The latter is indirectly supported by the actual existence of promiscuous compounds, known to belong to both the Hep and other classes, as in Table 12. They are not the ones contributing to the dilution of Hep-positive population in the right-hand plot of Figure 31 (when in several classes, compounds are counted as “blue” in “class versus remainder of antiviral” plots), but they are indirect evidence in favour of the possibility of promiscuous molecules(Figure 32).



CHEMBL1643 (Ribavirin)
Activity: *Hep, Inf, Pes, Sim*



CHEMBL1940452
Activity: *Hep, Pes, Sim*

Figure 32. Examples of promiscuous antivirals

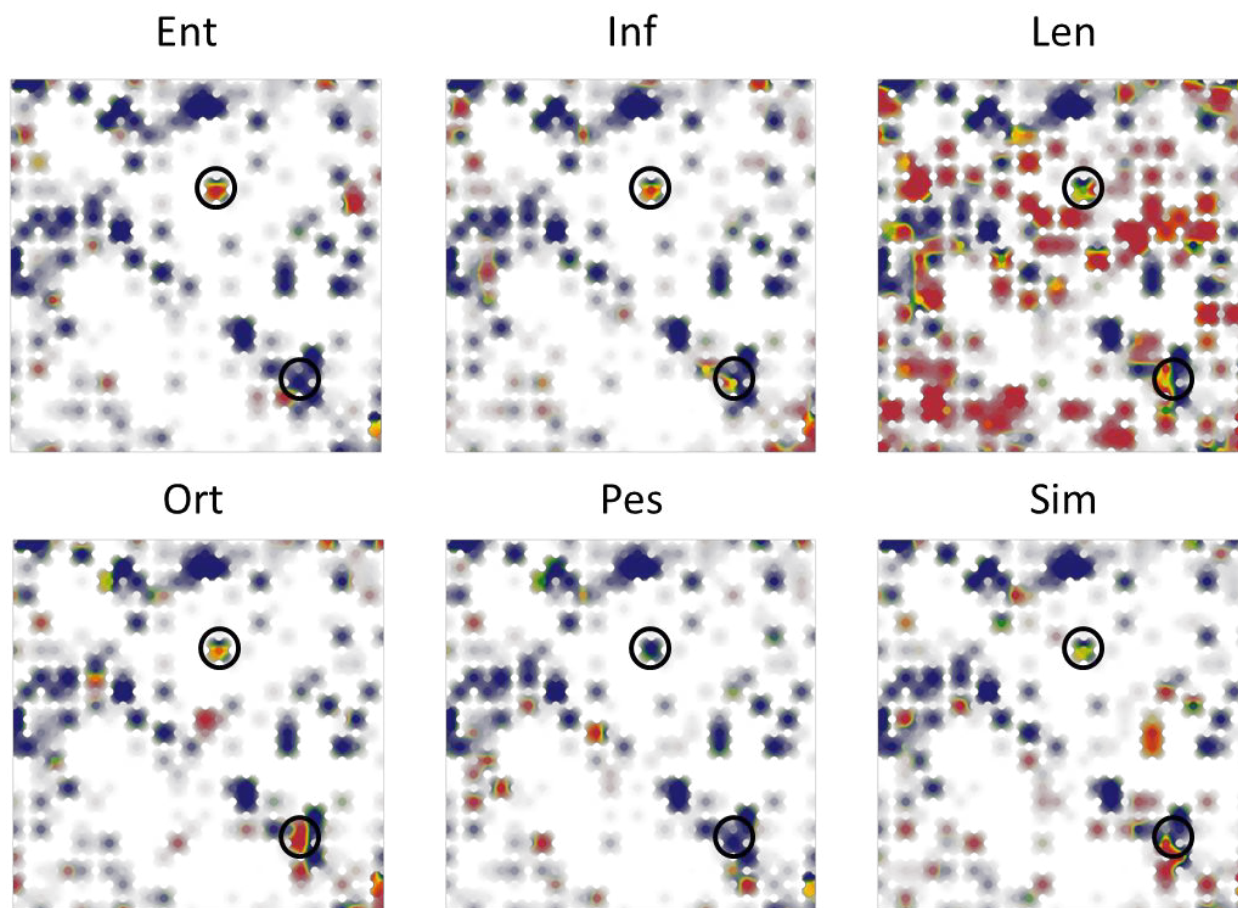


Figure 33. Pairwise classification landscapes built for Map 1 confronting hepacivirus positives (dominant in blue areas) with positives of the six other antiviral classes, respectively. Encircled zones correspond to two of the problem spots highlighted in Figure 31, with the third – the southeast corner – seemingly attracting positives of all the classes.

More details can be provided by constructing specific “class versus class” landscapes, by contrast to the above “one class versus remaining antivirals” (Figure 33).

Plots of Hep-positives (set as class “2”, blue) versus the positives of the six other classes (each in the role of class “1”, red) may reveal which are the other classes that overlap with the problematic areas in Figure 31. If a class does not interfere, the areas should be Hep-dominated (blue): Pes-positives are absent from the both encircled areas, whereas Ent-positives are absent from the “south-east” area only. The corner zone, not encircled, seems to have little specificity, and harbours structures of all the seven classes – probably a “garbage” area receiving compounds that are not closely approached by the manifold.

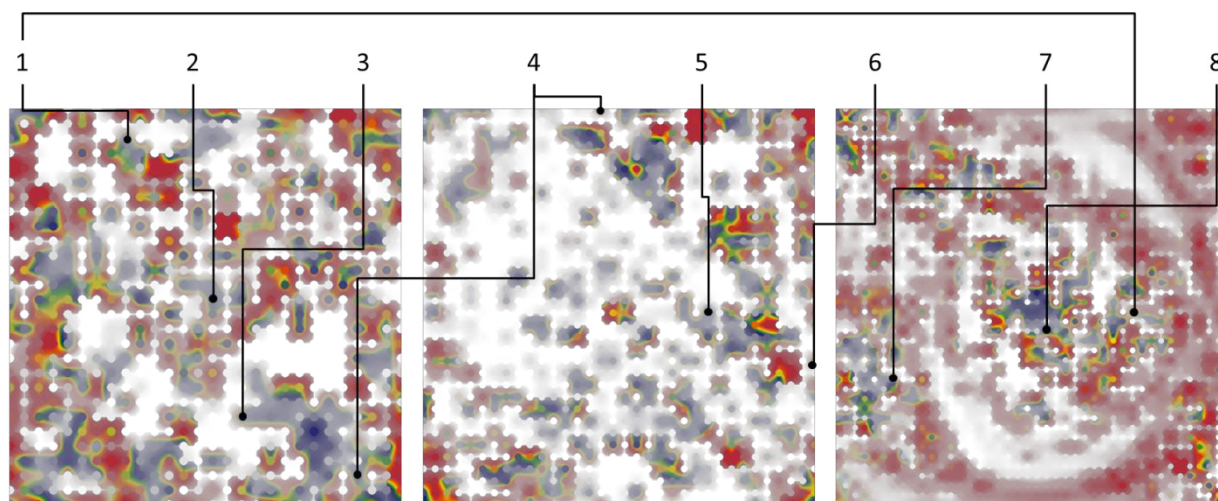


Figure 34. Classification landscapes of antiviral compounds, all classes confounded and labelled “2” (blue) versus non-antiviral ChEMBL molecules (status “0” in the original classification, here acting as class “1”, in red). Highlighted nodes on the maps (#1,2,3 from left to right, as in Table 10) correspond to the single-node PRPs harbouring the PSMs shown in Table 12.

Eventually, even though ChEMBL compounds labelled as non-antivirals (status “0”) were never used in the map selection process, Figure 34 above clearly illustrates that the maps are nevertheless well able to distinguish these from the antiviral molecules. This is not a trivial result, for unlike the “universal” maps reported in a previous work, [142] the frame sets used for antiviral map building failed to include major drug-like categories such as GPCR binders. On low-resolution maps like map 1, these various “novel” chemotypes not covered by the frame set seem to collapse into a few very high-density nodes (7 nodes have cumulated responsibilities above 25 thousand compounds each). By contrast, in higher-resolution map 3, the non-antiviral compounds seem to map onto the zones that were left largely empty by the antiviral molecules.

5.3.2. Analysis of the privileged patterns

Every compound out of 1.2 M receives an RP which can be used to group them regardless of their activity status. However, we are only interested in the ones that can lead to structural features behind antiviral activity. Therefore, it was crucial to develop an evaluation parameter to select activity-saturated patterns for further consideration.

Compounds featuring a common responsibility pattern (*RP*) on a GTM, can be regarded as a “cluster”. If such “cluster” contains a significantly high percentage of molecules associated with a given activity class (with respect to the entire library), then the pattern defining the cluster (*RP*) is *privileged* [152,153] with respect to that activity class. Let $f_{act}(RP)$ represent the fraction of “active” compounds matching a given pattern *RP*, where “active” should be understood in the broad sense of molecule having a desired property, belonging to a given therapeutic class, having a special status. By contrast, let $f_{def}(RP)$ represent the default fraction of molecules, out of the entire collection under study and related to the pattern. A privileged pattern *RP* associated primarily with the “actives” will have $f_{act}(RP)/f_{def}(RP) \gg 1$. Therefore, map zones corresponding to the privileged patterns are saturated with active compounds.

Here, two distinct types of “privilege” will be defined. On the one hand, one may check whether a pattern is seen more often within all antiviral compounds (positives of the seven classes, plus other antivirals), with respect to the entire ChEMBL database. This Antiviral specificity score (*Asp*) can be thus defined as:

$$Asp(RP) = \frac{f_{antiviral}(RP)}{f_{CHEMBL}(RP)} \quad (21)$$

where $f_{antiviral}(RP)$ the pattern occurrence frequency within the antiviral compound set, while $f_{CHEMBL}(RP)$ is the default pattern occurrence frequency within the 1.2M ChEMBL compounds.

On the other hand, it is interesting to assess whether a pattern is privileged by compounds associated with a given antiviral class, with respect to its occurrence frequency among all antivirals. This class specificity, *Csp*, can be written as:

$$Csp(RP@C) = \frac{f_{positives\ of\ class\ C}(RP)}{f_{antiviral}(RP)} \quad (22)$$

with $f_{positives\ of\ class\ C}(RP)$ being the RP occurrence frequency within the subset of positives of the class C. A number of patterns and scaffolds were found to be prevalent with both the antiviral status in general and specific classes in particular. The most prominent Privileged Responsibility Patterns (PRP) were selected for an in-depth

discussion if it was seen to occur at least 20 times within the positives of some activity class, and both *Asp* and *Csp* (for at least one of the seven C) reached values of 10 or more.

Privileged Responsibility Pattern description is the key to an intuitive understanding of the chemical meaning of specific chemical space zones. The most relevant privileged responsibility patterns (PRPs), satisfying the empirical criteria shown before, refer to single map nodes, monopolizing 100% of the responsibility distribution of associated compounds. In this specific case, PRPs may be thought of as “privileged map nodes”. These nodes were highlighted in Figure 34, in the context of displaying the generic antiviral compound space by contrast to the rest of the ChEMBL. The key PRPs are both specific to antivirals versus non-antiviral compounds, and, furthermore, specific to some antiviral classes as by contrast to the remainder of antiviral chemical space. Therefore, they consistently fall within blue, antiviral chemical space zones – but do not represent maximal density areas.

It is important to define the key structural features causing compounds to map by a same PRP, and thus providing – by extrapolation – their antiviral activity. These key features were termed Privileged Structural Motifs (PSMs) (Figure 35). Common structural features of “privileged” compounds associated with each PRP were determined by visual inspection. Note that a same PSM might be independently discovered as underlying structural motifs of PRPs on different maps (PSM number, same as in Figure 34)

#	Class	Privilege Structural Motif	Representative molecules
1	<i>Pes</i>	<p>Various linkers</p> <p>Various substituents</p>	
2	<i>Ort</i>	<p>or</p>	
3	<i>Sim</i>		
4	<i>Sim</i>		
5	<i>Sim</i>	<p>[C;N]</p> <p>(L1-2)</p>	
6	<i>Sim</i>	<p>Nucleoside-like heterocycle</p> <p>various linker chains</p>	
7	<i>Ort</i>		

8 *Inf*

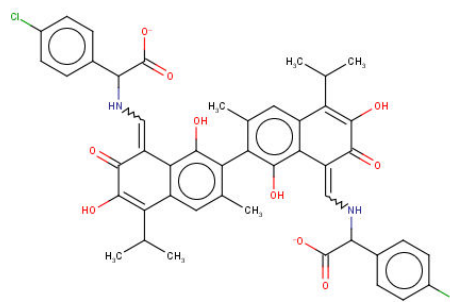
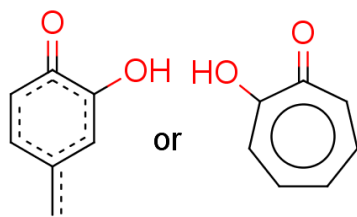


Figure 35. Main privileged antiviral structural motifs resulted from the analysis of responsibility-based patterns. Location of the structures on the maps is shown in Figure 34

PSM 1 was selected because it is privileged by the pestivirus class, and was convergently discovered in PRPs of both maps 1 and 3. This node harbours 36 of the 412 pestivirus-positives, which brings its Csp (Pes) value to 17. However, it is dominated by 120 anti-HCV compounds which are believed to be HCV non-structural proteins inhibitors [154,155]. Since, however, the pool of hepacivirus-positives is intrinsically larger (>5300), the 120 compounds only account for a hepacivirus-class Csp of 4.5. Note that 34 compounds of the 36 Pes-positives are actually labelled as both Hep and Pes. Most compounds (110 out of 120), such as Daclatasvir, are imidazolylpyrrolidines.

As this example clearly shows, privileged status of a RP with respect to a class does not mean that the given class is necessarily the best represented within that RP – it means that a relative majority of compounds from that class match the given RP. Compounds of other classes may dominate that RP – yet, if they represent less significant fractions of their respective classes, this RP may be less privileged with respect to the latter. The example also suggests that the arbitrary factor of 10 chosen to pick the most “extreme” cases of privileged patterns for discussion is far too restrictive: useful insight may be gained from patterns at lower values. Note that for Hep and Len, Csp scores of 10 are impossible due to their high occurrence in the antiviral dataset. Even if a given RP would exclusively occur within one of these classes, $f_{\text{positive}} = (X \text{ occurrences} / F \text{ class members})$ reported to antiviral = $(\text{same } X \text{ occurrences} / A \text{ antivirals})$ cannot exceed A/F , a ratio well below 10 for either of Hep and Len. The above-mentioned Csp (Hep) of 4.5 is virtually equal to the absolute maximum $A/F = 24629/5320 = 4.63$ achievable within this data collection. The absence in Figure 35 of PRPs specifically dedicated to the two main antiviral classes is not a problem – they were detected during the analysis, but were not picked for the present proof-of-concept discussion of PRPs as direct sources of privileged structural motifs.

PSM 2, privileged by the Orthohepadnavirus class with a Csp exceeding 30, is harbouring a structurally diverse set of compounds possessing different activities, such as anti-measles [156] and anti-HBV (HBsAg inhibitors) [157]. The majority of them are variations of either of the two distinct but similar scaffolds shown in Figure 36. Hence, this PSM regroups a pair of similar scaffolds that are allegedly interchangeable options in antiviral compound design –as the knowledge extracted by Generative Topographic Mapping seems to suggest.

Both **PSMs 3** and **4** (privileged by simplexvirus) consists of nucleoside-based analogues with broad-spectrum of antiviral activity. This is a case where the highlighted “structural motif” coincides with the classical definition of privileged scaffolds. The first PSM is mainly seen in anti-HSV [158] and anti-HIV [159] compounds, while the second, mainly anti-HSV-oriented [160,161] includes popular drugs Ganciclovir and Aciclovir.

PSM 5 (privileged by simplexvirus) comprises antivirals with various polyheterocyclic systems [162]. This pattern regroups a series of very close but distinct scaffolds, differing in terms of ring size (5 versus six-membered) and the positions of aromatic N atoms on the otherwise conserved scaffold graph. This example shows that responsibility patterns are able to spontaneously regroup closely related scaffolds.

PSM 6 regroups various anti-simplexvirus nucleotide mimics, with various heterocycles (including, but not restricted to, the natural purines and pyrimidines) linked via a linear chain (mainly hydrophobic, occasionally including an ether group) to a phosphate group. Clearly, on one hand **PSM 6** cannot be reduced to any single scaffold, while, on the other, it is not solely defined by the scaffold. Representatives of this class also display broad spectrum of activity, particularly against HIV [163] and herpesviruses [164].

PSM 7 (Orthohepadnavirus) translated into a very homogeneous family of steroid compounds, such as caudatin and its derivatives, which originally come from natural sources and were found effective against HBV [165]. It is noteworthy that in this case the actual definition of the privileged motif is very precise: more specific than the mere scaffold structure. Not only the scaffold per se, but also some of its “ornaments” appear to be conserved throughout the group. Note that the retrieved pattern is chiral, albeit chirality is ignored by the used molecular descriptors. This should not be interpreted as some prediction of the required chirality, but simply as an observation that the current motif systematically appears under this single stereochemistry in the database, which is

not surprising within a series of chemically modified natural products. Would ChEMBL have contained different stereoisomers of this moiety, those would have been mapped onto the same node, and, if not listed amongst known antivirals, would have “eroded” the privileged status of the pattern. This clearly shows (a) the intrinsic limitation of any 2D descriptor-based analysis and (b) that a privileged status is often not a reflection of the intrinsic preference of the target for that moiety, but a mere bias due to absence of “negative” counterexamples featuring that pattern.

PSM 8 emerges as privileged of both Influenza A and Orthohepadnavirus classes and covers a rather diverse structural family, most of which (but not all) have in common the benzoquinone dimer highlighted in Figure 36, embedded in a large variety of chemical contexts. Counterexamples not featuring this dimer core contain a single quinone moiety or, alternatively, a tropolone core. Many of the species appear as negatively charged at physiological pH, either due to ionization of rather ubiquitous phenol groups, or due to the presence of sulfonate and carboxylate anions. Albeit this motif does not seem to allow any simple definition in terms of common scaffolds, regrouping these – putatively redox-active – quinone/polyphenols together does make perfect chemical sense. It is an example of a fuzzy but meaningful motif that could not have been highlighted as such by substructure mapping. The compounds display antiviral activity against HIV [166,167], HBV [168] and Influenza A [167].

Thus, the analysis of PRP-based compound clusters turned out to be a tool of high versatility, because it does not rely on any preconception on the nature of the structural motif to look for. Sometimes, the PSM found to characterize the given subset of antivirals actually happens to coincide with the presence of a privileged antiviral scaffold – nucleosides, notably. However, in some cases the actual motif may be more finely tuned than simple scaffold presence – the privileged structure may be a specifically substituted scaffold, not any occurrence thereof. By contrast, sometimes the common characteristic of an antiviral compound subset may be too fuzzy to pinpoint in terms of specific substructures, all while making nevertheless perfect chemical sense, as was the case of the rather diverse phenol/quinone species, or the series of rather diverse nucleotide mimics. Note that some PSM are being specifically “discovered” by several maps, each independently allotting a node for harbouring broadly the same subset of structurally related compounds. By contrast, others are specifically highlighted by only one of the three maps, which are thus able to provide complementary perspective overviews of the antiviral space.

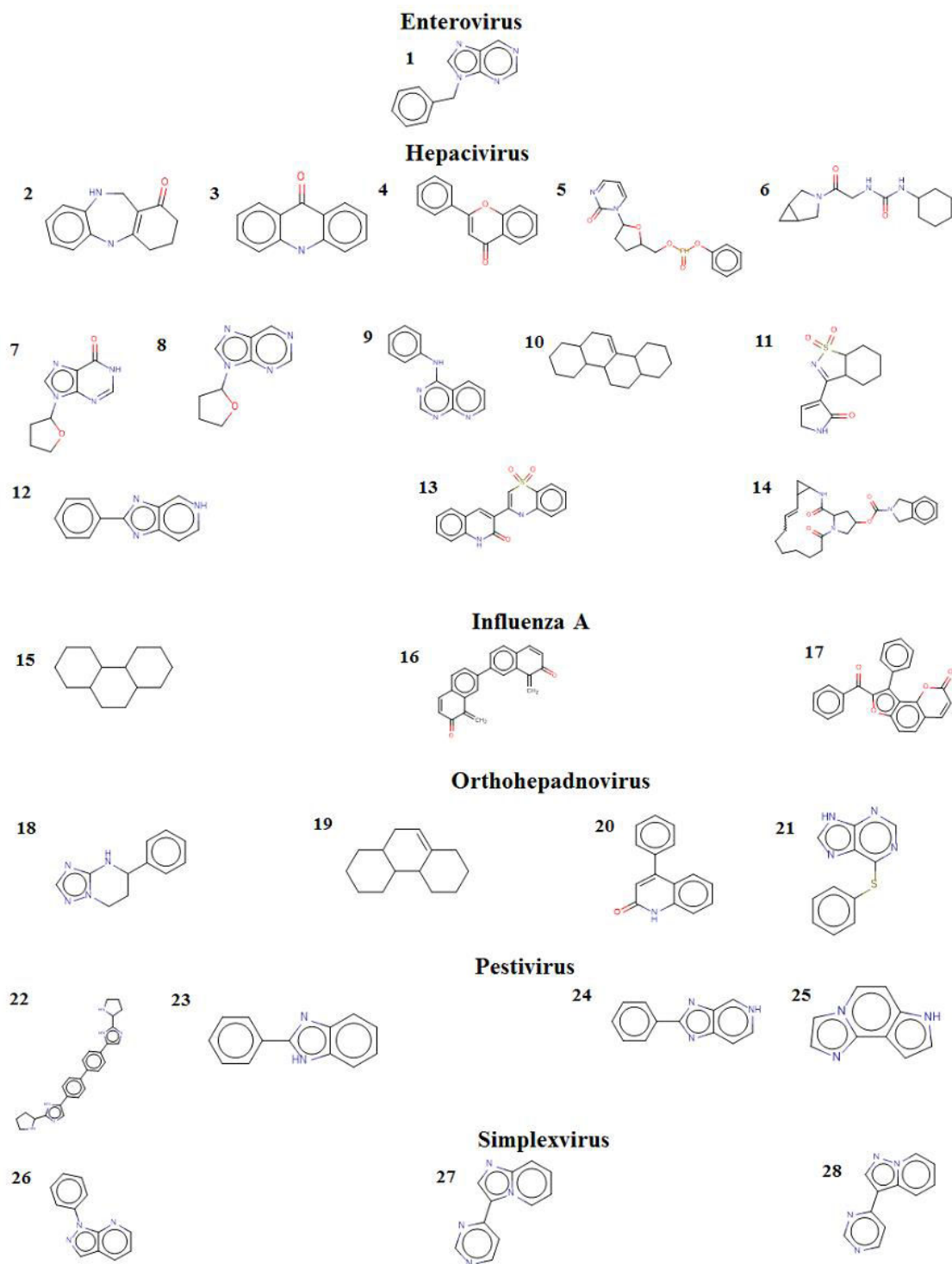


Figure 36. Scaffolds for antiviral compounds, extracted for each antiviral class by Scaffold Hunter

It is interesting to note that virtually all highlighted patterns have strong relatedness to classes of natural compounds – peptides, nucleosides, sterols and polyphenols. Natural compounds or derivatives thereof appear to be privileged in antiviral research,

perhaps more than in other “more rational” branches of drug design. This trend is spontaneously highlighted by the GTM-driven analysis.

Since GTM is a relatively new technique in the field of chemoinformatics, it was important to compare it to something already well-established. In our case GTM provides visualization, activity class prediction and chemical space analysis of the dataset. However, comparing every aspect of GTM to already existing tools would be excessive, therefore we decided to compare its “analytical” powers to the scaffold approach which has already been used for some time. Scaffold hunter software [169] was used to define compounds chemical class by determining the most common substructures within the major activity classes – except the set of lentivirus-positives, which is too large and too diverse for the present purpose. This software organizes scaffolds in a tree-like hierarchy based on the inclusion relation, enabling navigation in the associated chemical space in an intuitive way. The criteria of being “privileged” was applied for the best scaffold selection in the same manner it was applied to RPs.

When the privileged status of “naked” scaffolds was assessed, it was seen that among the 28 checked scaffolds (Figure 36), only 5 correspond to the criteria of being “privileged”, particularly scaffolds # 16, 17, 22, 25 and 28. Some of these were already discussed, because they are present within the responsibility-driven compound clusters in Figure 35. Even scaffolds that co-define fuzzier PSMs in combination with different, related substructures may nevertheless score high Asp and Csp values – classical analysis would have highlighted them as privileged, whilst in fact they are only peculiar “incarnations” of a broader motif as highlighted previously. Such examples include scaffold #16, a frequent representative of PSM 8, coexisting next other various cores of hydroxylated quinone or tropolone type. Similarly, scaffold #22 is one peculiar substructure appearing in PSM 1, and the same applies for scaffold #28 with respect to PSM 5. Scaffold #25 compounds are active against BVDV-1 [170] while scaffold #17 is privilegedly encountered in positives of the Influenza A class.

5.4 Activity prediction of antiviral CADD compounds

Note that these maps may allow for an even deeper interpretation of external compounds projection results. After their responsibility pattern is defined, training set antivirals which share the same RP are individually examined.–This involves in-depth verification of activity type (target virus) and bioassay description sources – from any

source of information, including of course the already preprocessed ChEMBL. Note that, by default, external compounds are expected to be similar to their training set counterparts sharing the same RP – if, for ever reason, this is not the case, it indicates that the external compound is too ‘exotic’ with respect to any of the originally used frame set molecules and hence outside of the AD of the method. If the specific experimental data gathered for the training RP members seems to converge towards a coherent therapeutic indication, then this is proof in favor of the working hypotheses that the given RP might be associated to that therapeutic indication. This hypothesis then automatically extends to the RP-matching external compounds. The advantage here is that the GTM approach allows to ‘focus’ on RP-specific compound subsets, for which mining of antiviral information may be pursued manually, in much greater detail, without the constraints of the automatized record extraction used so far, *i.e.* a *local* coloring of the map in much subtler “nuances” that the 7 viral categories can be envisaged. This is particularly useful for identification of source publication with detailed bioassay description, allowing researchers to evaluate antiviral data reliability before making decision about using particular bioassay or its specifications (i. e. target virus, host cells, activity measurement technique) to test new compounds.

In this study two compounds (**10** and **24**) from CADD part were projected (Figure 37) on 3 developed GTMs (Table 12).

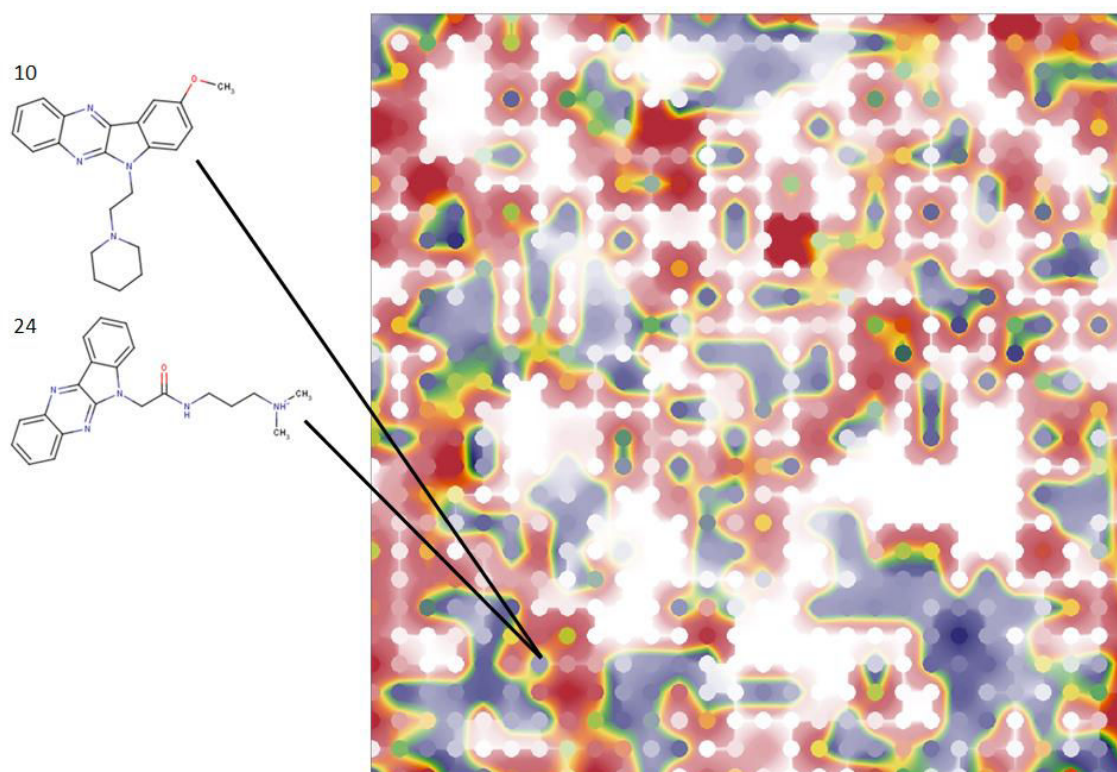


Figure 37. CADD hit compounds projection on the map 1

None of them matched any privileged responsibility patterns outlined so far. Neither **10** nor **24** should be active against any major antiviral *Genus*, according to the maps. However, training compounds of the same responsibility pattern could be divided in three groups. First group contains indolequinoxaline compounds which were synthesized and tested against VSV by the PCI laboratory [8,9] and became the part of training set for antiviral QSAR models described in Part 4. The second group were indolquinoxalines synthesized outside PCI which were found active against *Vaccinia virus* [171] and *Cytomegalovirus* [172]. The third group consists of triazolotriazinoindoles with the potent activity against *Mammalian orthoreovirus* [173]. These findings mean that GTMs were able to match projected compounds with previously synthesized ones from the same lab, as well as, with compounds of the same class synthesized elsewhere and successfully tested against the same pathogen (*Vaccinia virus*) and another double-stranded DNA virus (CMV). But finding similarity between indolequinaxolines and triazolotriazinoindoles is of the most importance since they belong to different families but can potentially have similar mechanism of action and activity spectrum. In terms of molecular structure, above-mentioned compounds consist of monosubstituted planar aromatic heterocyclic system of approximately the same size (Figure 38 green outline). This structural feature ensures intercalating capacity of indolequinaxolines and can possibly do the same for triazolotriazinoindoles, making the second ones DNA binders as well. As for activity, *Mammalian orthoreovirus* is a double-stranded RNA virus, which make it a suitable target for CADD 'hit' compounds



Figure 38. Comparison of indolequinaxoline (cmpd. 10) and triazolotriazinoindole (CHEMBL2296704). Similar fragment is outlined in green, dissimilar - in red.

Residues in indolequinaxolines and triazolotriazinoindoles are quite different: the triazolotriazinoindoles lack positively-ionizable group, whereas it has a non-fused aromatic ring (Figure 38 red outline). Presented triazolotriazinoindoles have a larger DNA binding site size compared to indolequinaxolines which according to some researchers

[174] contributes to increase in anti-tumor activity and, possibly toxicity. Therefore, an absence of aromatic ring in the side chain of indolequinaxoline analogues of triazolotriazinoindoles can be a sign of their lower toxicity; this is an assumption which could be put to test by an anti-MRV study of above-mentioned indolequinaxolines.

5.5 Conclusions

The first important conclusion is that so-far available public data is largely insufficient in order to allow a rigorous buildup of an actual structure–antiviral activity profile. Experimental information is sparse, as no compound has been systematically tested against all the virus strains. Therefore, data fusion of individual assay results, in order to assign generic antiviral class-based membership labels is, so far, the only way we found to extract statistically exploitable training sets supporting the attempted analysis of antiviral chemical space.

To sum up the analysis, this work represents an audit of antiviral structure-activity data in the public database ChEMBL, using chemoinformatics tools and in particular aimed at showing how the rather recent technique of Generative Topographic Mapping (GTM) may be used to rationally render, intuitively visualize, model and predict antiviral activities as a function of compound structure. More precisely, targeted goals were to:

- Curate and standardize structure-antiviral activity in ChEMBL,
- Provide an association of individual structures with seven broad virus classes, transcending the numerous and diverse antiviral test protocols, thus building large and robust structure-class training sets that are perfectly suited for categorical model building
- Build dedicated GTMs that optimally discriminate between the above-mentioned classes, and to use these for visualization of the antiviral chemical space in the context of the entire ChEMBL compound collection, and of the specific space zones allotted to the specified antiviral classes.
- Use generated maps to focus attention on specific responsibility patterns – corresponding to specific locations on the map – that appear as “privileged” by certain antiviral classes.
- Understand how these responsibility patterns relate to the structural features of molecules seen to cluster together, and compare insights that can be gained from privileged responsibility patterns to the classical scaffold-based “privileged structure” analysis.

Despite the intrinsic uncertainty of used class labels (a “negative” may be wrongly assigned, because so far not tested on that virus class), GTMs successfully separated positives from negatives, with 3-fold cross-validated balanced accuracy scores of 0.8-0.9. External validation – challenging the maps to detect antiviral compounds outside of the ChEMBL database, and not used at map growing stage – was a partial success, especially with respect to Influenza A compounds (20 out of 21 were recognized as such, after mapping).

Visually, the separation into classes, each preferentially mapping to other areas on the map, can be clearly observed, which may enable particular observations concerning specific map zones. For example, it is straightforward to visualize the relative positioning of the positives associated with any two antiviral classes, observing their potential overlap zones where “promiscuous” compounds may reside. This may open

perspectives for antiviral compound repurposing, if the compound is seen to reside in an overlap zone involving its so-far targeted virus class and another, yet unassessed virus class.

Detection of responsibility patterns that are “privileged” by any antiviral class (in the sense of occurring more often than expected on a random basis within the associate “positive” compounds) was proven to represent a powerful generalization of the search for “privileged structures”, a paradigm in medicinal chemistry. Establishing the “privileged” status of a structural motif is a simple statistical exercise – however, a medicinal chemist is facing a virtual infinity of possible motifs (substructures, connected or disjointed graphs, pharmacophore patterns etc.) for which the privileged status should be assessed. Typically, they focus their attention on scaffolds – (poly)cyclic cores, or any other definition that is convenient for this rather fuzzy concept. The use of GTM technology, provides an answer to the key question “what is the nature of the privileged structural features?” The structural motifs shared by all the compounds represented by a same PRP are likely to be an excellent choice to systematize the essential characteristics of antiviral compounds. A PRP can be based not on one or several – different, yet quite similar, from the antiviral perspective “interchangeable” scaffolds, an aspect that would have been difficult to grasp when looking at each of those individual scaffolds. A PRP may also turn out to be more specific than the bare scaffold. Therefore, responsibility pattern analysis is a powerful application of GTM technology, able to spontaneously adjust to the correct “resolution” needed:

- at scaffold level. In some cases, privileged responsibility patterns are seen to gravitate around a common scaffold, which could have been picked as “privileged” by a classical analysis.
- coarser than scaffold level (such as the large and diverse family of polyphenols/quinones/tropolones, where the common trait seems to be the redox-competent functional groups rather than any specific scaffold).
- finer than scaffold level, the virus group turns out to privilege not the scaffold per se, but a specifically substituted scaffold, as exemplified by the substituted steroid core in Figure 35. In this case, the scaffold alone might not even be recognized as privileged, and the insight would have been lost in classical analysis.

The nonlinear nature of GTM models coupled to the evolutionary optimization including the selection of best suited molecular descriptor schemes, bound to capture relevant structural information allows to automatically tune in to the best resolution level needed to capture privileged structural characteristics in general, rather than predefined scaffolds that may or may not match the trend present in experimental data. Some privileged structural motifs were being reproducibly highlighted by several of the maps, each independently allotting a node for harboring roughly the same subset of structurally related compounds. By contrast, other chemical features are specifically highlighted by only one of the three considered antiviral maps. These different mapping schemes, based on different molecular descriptor sets, are thus partly convergent and partly complementary in terms of the light they shed on the antiviral space.

GENERAL CONCLUSIONS

1. An ensemble of the modeling tools including structure filters, pharmacophore and QSAR models developed in this work, as well as assessment of side effects and some ADME/Tox properties with commercial software has been used to perform screening of the large database of some 3M compounds. Virtual screening resulted in 55 compounds which then have been synthesized and tested experimentally. Biological experiments revealed substantial antiviral activity of two compounds screened against *Vaccinia virus*. These compounds displayed low toxicity at activity doses and proved to be DNA-binding ligands, which supports our suggestion of nucleic acid intercalation as the supposed mechanism of antiviral activity.
2. A comprehensive Antiviral database has been created using information about 24 629 compounds annotated with activity against more than 100 virus species extracted from the ChEMBL database. These data were curated and annotated with Virus *Genus*. This taxonomic rank is quite universal to group viruses with the similar basic characteristics (e. g. genome size, virion shape) and pathogenicity while being specific enough to differentiate between viruses with rather different protein composition.
3. Generative Topographic Mapping was used for analysis, visualization and activity class prediction of compounds from antiviral database. Three top fitness maps were obtained from an evolutionary process browsing through the space of possible GTM setups. Two-class GTM-Based classification models performs reasonable activity prediction with respect to each major *Genus* in 3-fold cross-validation (Balanced Accuracy varies from 0.77 to 0.91) and on the external test set (11 out of 19 compounds were predicted correctly).
4. Data visualization provided a simple notion of map regions enriched with active compounds. These regions can be a subject to further analysis and extraction of structural features appearing to be covariant with certain biological activities. Analysis of compounds in these zones allowed revealing 8 privileged structural motifs ensuring particular antiviral activity. Their structural features varied, from very detailed substructure (PSM7, anti-orthohepadnavirus) and classical scaffolds (PSM4, anti-simplexvirus) to a set of interchangeable scaffolds (PSM5, anti-simplexvirus) and even fuzzy common fragments (PSM8, anti-Influenza A).

5. GTM-derived Privileged Responsibility Pattern approach for chemical space analysis was compared to classic scaffold analysis. Scaffolds were derived from the class-specific subsets of the antiviral database, and then criteria of being “privileged” were applied to them. The scaffold approach yielded only 5 “privileged” structures: 3 out of them were particular cases of PSMs. This shows that PRPs are a more general way to approach the problem of “privileged” structural motifs, because they encompass privileged scaffolds as particular cases.
6. The two experimentally confirmed virtual screening hits were projected on maps and found in the area occupied by antiviral database compounds of a same structural class (indolequinaxolines), some of which are active against *Vaccinia virus* as well. Moreover, projected compounds have the same pattern as structurally similar triazolotriazinoindoles active against double-stranded RNA virus (*Mammalian orthoreovirus*). This allows assuming hit compounds may be active against MRV with a high chance of the positive outcome, which certainly requires an experimental confirmation.
7. QSPR model for predicting aqueous solubility in the temperature range 4-97 °C was developed using Random Forest method. Using k_j parameter allowed taking into account how particular compounds solubility is susceptible to the temperature change. Models were used to assess virtual screening hit candidates' solubility. Sample preparation for experimental part of CADD revealed that most compounds turned out to be soluble enough for biological testing.

REFERENCES

1. Shrestha SS, Swerdlow DL, Borse RH, Prabhu VS, Finelli L, Atkins CY, Owusu-Eduesei K., Bell B, Mead PS, Biggerstaff M, Brammer L, Davidson H, Jernigan D, Jhung MA, Kamimoto LA, Merlin TL, Nowell M, Redd SC, Reed C, Schuchat A, Meltzer MI. (2011) Estimating the Burden of 2009 Pandemic Influenza A (H1N1) in the United States (April 2009-April 2010). *Clin. Infect. Dis.*, 52 Suppl. 1: S75-82.
2. (2016). Ebola situation report, World Health Organization.
3. Brenner B, Turner D, Oliveira M, Moisi D, Detorio M, Carobene M, Marlink RG, Schapiro J, Roger M, Wainberg MA. (2003). A V106M mutation in HIV-1 clade C viruses exposed to efavirenz confers cross-resistance to non-nucleoside reverse transcriptase inhibitors. *Aids* 17 (1): F1-F5
4. Stringfellow DA, Glasgow LA. (1972). Tilorone hydrochloride: an oral interferon-inducing agent. *Antimicrob. Agents Chemother.* 2 (2): 73-78.
5. De Clercq E (2013) Antivirals: past, present and future. *Biochem. Pharmacol.* 85 (6): 727-744.
6. De Clercq E, Li G. (2016) Approved Antiviral Drugs over the Past 50 Years. *Clin. Microbiol. Rev.* 29 (3): 695-747.
7. Katz E, Margalith E, Winer B. (1975) Inhibition of Herpesvirus Deoxyribonucleic Acid and Protein Synthesis by Tilorone Hydrochloride. *Antimicrob. Agents Chemother.* 9 (1): 189-195.
8. Shibinskaya MO, Lyakhov SA, Mazepa AV, Andronati SA, Turov AV, Zholobak NM, Spivak NY. (2010) Synthesis, cytotoxicity, antiviral activity and interferon inducing ability of 6- (2-aminoethyl)-6H-indolo [2,3-b]quinoxalines. *Eur. J. Med. Chem.*, 45 (3): 1237-1243.
9. Shibinskaya MO, Karpenko AS, Lyakhov SA, Andronati SA, Zholobak NM, Spivak NY, Samochina NA, Shafran LM, Zubritskye MJ, Galat VF (2011) Synthesis and biological activity of 7H-benzo [4,5]indolo [2,3-b]-quinoxaline derivatives. *Eur. J. Med. Chem.*, 46 (2): 794-798.
10. Dimmock NJ, Easton AJ, Leppard KN. (2007). Introduction to modern virology, Blackwell Publishing.
11. Bukrinskaya AG. (1986) *Virusologia, Medicina*
12. Haim H, Steiner I, Panet A. (2007). Time frames for neutralization during the human immunodeficiency virus type 1 entry phase, as monitored in synchronously infected cell cultures. *J. Virol.* 81 (7): 3525-3534.
13. Moss B. (2012) Poxvirus cell entry: how many proteins does it take? *Viruses* 4 (5): 688-707.
14. Campadelli-Fiume G, Amasio M, Avitabile E, Cerretani A, Forghieri C, Gianni T, Menotti L. (2007). The multipartite system that mediates entry of herpes simplex virus into the cell. *Rev. Med. Virol.* 17 (5): 313-326.
15. Sebestyén MG, Budker VG, Budker T, Subbotin VM, Zhang G, Monahan SD, Lewis DL, Wong SC, Hagstrom JE, Wolff JA. (2006). Mechanism of plasmid delivery by hydrodynamic tail vein injection. I. Hepatocyte uptake of various molecules. *J. Gene. Med.* 8 (7): 852-873.
16. Helle F, Dubuisson J. (2008). Hepatitis C virus entry into host cells. *J. Cell. Mol. Life Sci.* 65 (1): 100-112.
17. ViralZone: www.expasy.org/viralzone, SIB Swiss Institute of Bioinformatics
18. Furukawa T, Muraki Y, Noda T, Takashita E, Sho R, Sugawara K, Matsuzaki Y, Shimotai Y, Hongo S. (2011). Role of the CM2 protein in the influenza C virus replication cycle. *J. Virol.* 85 (3): 1322-1329
19. Crick F. (1970). Central dogma of molecular biology. *Nature* 227 (5258): 561-563

20. (1999). Assessment of Future Scientific Needs for Live Variola Virus, Institute of Medicine (US) Committee on the Assessment of Future Scientific Needs for Live Variola Virus.
21. Garfinkel MS, Katze MG. (1993). Translational control by influenza virus. Selective translation is mediated by sequences within the viral mRNA 5'-untranslated region. *J. Biol. Chem.* 268 (30): 22223-22226.
22. Finke S, Conzelmann KK. (2005). Replication strategies of rabies virus. *Virus Res.* 111 (2): 120-131.
23. Horspool D. https://en.wikipedia.org/wiki/File:Central_Dogma_of_Molecular_Biochemistry_with_Enzymes.jpg
24. Sundquist WI, Krausslich HG. (2012). HIV-1 assembly, budding, and maturation. *Cold Spring Harb Perspect Med* 2 (7): 1-25.
25. Domingo E, Holland JJ. (1997). RNA virus mutations and fitness for survival. *Annu. Rev. Microbiol.* 51: 151-178.
26. Drake JW. (1993). Rates of spontaneous mutation among RNA viruses. *Proc. Natl. Acad. Sci. USA.* 90 (9): 4171-4175.
27. Greene EA, Codomo CA, Taylor NE, Henikoff JG, Till BJ, Reynolds SH, Enns L C, Burtner C, Johnson JE, Odden AR, Comai L, Henikoff S. (2003). Spectrum of chemically induced mutations from a large-scale reverse-genetic screen in *Arabidopsis*. *Genetics* 164 (2): 731-740.
28. Carr JG (1948) Chemically induced mutation. *Br. J. Cancer.* 2 (2): 132-134.
29. Sanjuan R, Nebot MR, Chirico N, Mansky LM, Belshaw R. (2010). Viral mutation rates. *J. Virol.* 84 (19): 9733-9748.
30. Ribeiro RM, Qin L, Chavez LL, Li D, Self SG, Perelson AS. (2010). Estimation of the initial viral growth rate and basic reproductive number during acute HIV-1 infection. *J. Virol.* 84 (12): 6096-6102.
31. World Health Organisation (2011) Chapter 1: Smallpox: eradicating an ancient scourge, from Bugs, drugs and smoke: stories from public health, p. 3-21 http://www.who.int/about/history/publications/public_health_stories/en/
32. Uldrick TS, Whitby D. (2011). Update on KSHV epidemiology, Kaposi Sarcoma pathogenesis, and treatment of Kaposi Sarcoma. *Cancer. Lett.* 305 (2): 150-162.
33. Parkin DM (2006). The global health burden of infection-associated cancers in the year 2002. *Int. J. Cancer.* 118 (12): 3030-3044.
34. Ruuskanen OE, Lahti E, Jennings, L. C. Murdoch, D. R. (2011). Viral pneumonia. *Lancet* 377 (9773): 1264-1275.
35. Petrovsky NJ, Aguilar C. (2004). Vaccine adjuvants: current state and future trends. *Immunol. Cell Biol.* 82 (5): 488-496.
36. Badgett MR, Auer A, Carmichael LE, Parrish CR, Bull JJ. (2002). Evolutionary dynamics of viral attenuation. *J. Virol.* 76 (20): 10524-10529.
37. Bayer ME, Blumberg BS, Werner B. (1968). Particles associated with Australia antigen in the sera of patients with leukaemia, Down's Syndrome and hepatitis. *Nature* 218 (5146): 1057-1059.
38. World Health Organization (2014) Polio vaccines: WHO position paper *Wkly. Epidemiol. Rec.* 89 (9): 73-92.
39. World Health Organization (2010) Rabies vaccines: WHO position paper--recommendations. *Wkly. Epidemiol. Rec.* 28 (44): 7140-7142.
40. World Health Organization (2012). Vaccines against influenza WHO position paper *Wkly. Epidemiol. Rec.* 87 (47): 461-476.
41. Fiore AE, Bridges CB, Cox NJ (2009). Seasonal influenza vaccines. *Curr. Top. Microbiol. Immunol.* 333: 43-82.

42. Sekaly R. P. (2008). The failed HIV Merck vaccine study: a step back or a launching point for future vaccine development? *J. Exp. Med.* 205 (1): 7-12.
43. World Health Organization (2013). Model list of Essential Medicines, 18th edition: 1-43.
44. Pillai B, Kannan KK, Hosur MV. (2001). 1.9 Å X-ray study shows closed flap conformation in crystals of tethered HIV-1 PR. *Proteins* 43 (1): 57-64.
45. Varghese JN, Colman PM. (1991). Three-dimensional structure of the neuraminidase of influenza virus A/Tokyo/3/67 at 2.2 Å resolution. *J. Mol. Biol.* 221 (2): 473-86.
46. Fiore AE, Fry A, Shay D, Gubareva L, Bresee JS, Uyeki TM. (2011). Antiviral agents for the treatment and chemoprophylaxis of influenza --- recommendations of the Advisory Committee on Immunization Practices (ACIP). *MMWR Recomm. Rep.* 60 (1): 1-24.
47. De Andrea M, Ravera R, Gioia D, Gariglio M, Landolfo S. (2002). The interferon system: an overview. *Eur. J. Paediatr. Neurol.* 6 Suppl A: A41-46, discussion A55-58.
48. Samuel CE. (1979). Mechanism of interferon action: phosphorylation of protein synthesis initiation factor eIF-2 in interferon-treated human cells by a ribosome-associated kinase processing site specificity similar to hemin-regulated rabbit reticulocyte kinase. *Proc. Natl. Acad. Sci. USA* 76 (2): 600-604.
49. Zaliska, O., Piniashko, O, Maksymovych N, Sickhoriz O, Tolubaiev V. (2015). Pharmaceutical system in Ukraine: current and prospective issues. *J. Health Pol. Out. Res.* 2: 89-94.
50. Katz E, Margalith E, Winer B. (1976). The effect of tilorone hydrochloride on the growth of several animal viruses in tissue cultures. *J. Gen. Virol.* 31 (1): 125-129.
51. Steindl TM, Crump CE, Hayden FG, Langer T. (2005). Pharmacophore modeling, docking, and principal component analysis based clustering: combined computer-assisted approaches to identify new inhibitors of the human rhinovirus coat protein. *J. Med. Chem.* 48 (20): 6250-6260.
52. Gao L, Zu M, Wu S, Liu AL, Du GH. (2011). 3D QSAR and docking study of flavone derivatives as potent inhibitors of influenza H1N1 virus neuraminidase. *Bioorg. Med. Chem. Lett.* 21 (19): 5964-5970.
53. Zhou Z, Khaliq M, Suk JE, Patkar C, Li L, Kuhn RJ, Post CB. (2008). Antiviral compounds discovered by virtual screening of small-molecule libraries against dengue virus E protein. *ACS Chem. Biol.* 3 (12): 765-775.
54. Cheng LS, Amaro RE, Xu D, Li WW, Arzberger PW, McCammon JA (2008). Ensemble-based virtual screening reveals potential novel antiviral compounds for avian influenza neuraminidase. *J. Med. Chem.* 51 (13): 3878-3894.
55. Folkers G, van de Waterbeemd H, Lennernäs H, Artursson P, Mannhold R, Kubinyi H. (2003). Drug Bioavailability: Estimation of Solubility, Permeability, Absorption and Bioavailability (Methods and Principles in Medicinal Chemistry. Weinheim, Wiley.
56. (2015). Waiver of In Vivo Bioavailability and Bioequivalence Studies for Immediate-Release Solid Oral Dosage Forms Based on a Biopharmaceutics Classification System. Guidance for industry: 1-14.
57. Kholod YA, Muratov EN, Gorb LG, Hill FC, Artemenko AG, Kuz'min VE, Qasim M, Leszczynski J. (2009). Application of quantum chemical approximations to environmental problems: prediction of water solubility for nitro compounds. *Environ. Sci. Technol.* 43 (24): 9208-9215.
58. Yan A, Gasteiger J. (2003). Prediction of aqueous solubility of organic compounds based on a 3D structure representation. *J. Chem. Inf. Comput. Sci.* 43 (2): 429-434.

59. Butina D, Gola JM (2003). Modeling aqueous solubility. *J. Chem. Inf. Comput. Sci.* 43 (3): 837-841.
60. Tetko IV, Tanchuk VY, Kasheva TN, Villa AE (2001). Estimation of aqueous solubility of chemical compounds using E-state indices. *J. Chem. Inf. Comput. Sci.* 41 (6): 1488-1493.
61. Llinas A, Glen R, Goodman JM. (2008). Solubility challenge: can you predict solubilities of 32 molecules using a database of 100 reliable measurements? *J. Chem. Inf. Model.* 48 (7): 1289-1303.
62. Tetko IV, Bruneau P, Mewes HW, Rohrer DC, Poda GI. (2006). Can we estimate the accuracy of ADME-Tox predictions? *Drug. Discov. Today.* 11 (15-16): 700-707.
63. Handbook of pharmaceutical excipients (2009) Rowe RC, Sheskey PJ, Quinn ME, Eds, London; Chicago, APhA/Pharmaceutical Press
64. Tabaraki R, Khayamian T, Ensafi AA. (2006). Wavelet neural network modeling in QSPR for prediction of solubility of 25 anthraquinone dyes at different temperatures and pressures in supercritical carbon dioxide. *J. Mol. Graph. Model.* 25 (1): 46-54.
65. Leach AR, Gillet VJ. (2007). An Introduction to Chemoinformatics, Springer, 259
66. Gramatica P. (2013). On the development and validation of QSAR models. *Methods. Mol. Biol.* 930: 499-526.
67. Kuz'min VE, Artemenko AG, Muratov EN. (2008). Hierarchical QSAR technology based on the Simplex representation of molecular structure. *J. Comput. Aided. Mol. Des.* 22 (6-7): 403-421
68. Muratov EN, Artemenko AG, Varlamova EV, Polischuk PG, Lozitsky VP, Fedchuk AS, Lozitska RL, Gridina TL, Koroleva LS, Sil'nikov VN, Galabov AS, Makarov VA, Riabova OB, Wutzler P, Schmidtke M, Kuz'min V (2010). Per aspera ad astra: application of Simplex QSAR approach in antiviral research. *Fut. Med. Chem.* 2 (7): 1205-1226.
69. Laboratoire d'Infochimie UMR 7177 CNRS, Université de Strasbourg, 4, rue B. Pascal, Strasbourg 67000, France. <http://infochim.u-strasbg.fr/>.
70. Kuz'min VE, Artemenko AG, Polischuk PG, Muratov EN, Hromov AI, Liahovskiy AV, Andronati SA, Makan SY. (2005). Hierarchic system of QSAR models (1D-4D) on the base of simplex representation of molecular structure. *J. Mol. Model.* 11 (6): 457-467
71. Gaspar H, Marcou G, Horvath D, Arault A, Lozano S, Vayer P, Varnek A (2013) Generative topographic mapping-based classification models and their applicability domain: application to the biopharmaceutics Drug Disposition Classification System (BDDCS). *J. Chem. Inf. Model.* 53 (12): 3318-3325.
72. Kireeva N, Baskin I, Gaspar H, Horvath D, Marcou G, Varnek A. (2012). Generative Topographic Mapping (GTM): Universal Tool for Data Visualization, Structure-Activity Modeling and Dataset Comparison. *Mol. Inform.* 31 (3-4): 301-312.
73. Solov'ev VP, Varnek A, Wipff G. (2000). Modeling of ion complexation and extraction using substructural molecular fragments. *J. Chem. Inf. Comput. Sci.* 40 (3): 847-58.
74. Cormen TH, Leiserson CE, Rivest RL, Stein C. (2009). Introduction to Algorithms. London, MIT Press.
75. Kubinyi H., Ed. Mannhold R, Krogsgaard-Larsen P, Timmerman H. (1993). QSAR: Hansch Analysis and Related Approaches, Wiley, 252
76. Tropsha A, Gramatica P, Gombar VK (2003). The importance of being earnest: Validation is the absolute essential for successful application and interpretation of QSPR models. *QSAR Comb. Sci.* 22 (1): 69-77.

77. Breiman, L. (2001). Random forests. *Mach. Learn.* 45 (1): 5-32.
78. Polishchuk P, Muratov EN, Artemenko AG, Kolumbin OG, Muratov NN, Kuz'min VE. (2009). Application of random forest approach to QSAR prediction of aquatic toxicity. *J. Chem. Inf. Model.* 49: 2481-2488.
79. Breiman L. (2002). Manual On Setting Up, Using, And Understanding Random Forests V3.1.
80. Svetnik V, Liaw A, Tong C, Culberson JC, Sheridan RP, Feuston BP (2003). Random forest: A classification and regression tool for compound classification and QSAR modeling. *J. Chem. Inf. Comput. Sci.* 43 (6): 1947-1958.
81. Ognichenko LN, Kuz'min VE, Gorb L, Hill F, Artemenko A, Polishchuk PG, Leszczynski J. (2012). QSPR Prediction of Lipophilicity for Organic Compounds Using Random Forest Technique on the Basis of Simplex Representation of Molecular Structure. *Mol. Inf.* 31 (3-4): 273-280.
82. Kovdienko NA, Polishchuk PG, Muratov EN, Artemenko AG, Kuz'min VE, Gorb L, Hill F, Leszczynski J. (2010). Application of Random Forest and Multiple Linear Regression Techniques to QSPR Prediction of an Aqueous Solubility for Military Compounds. *Mol. Inf.* 29 (5): 394-406.
83. Breiman L, Friedman J, Olshen RA, Stone CJ. (1984). Classification and Regression Trees. Wadsworth, Belmont, CA: 368.
84. Klimenko K, Kuz'min V, Ognichenko LN, Gorb L, Shuckla M, Vinas N, Perkins E, Polishchuk P, Artemenko A, Leszczynski J. (2016). Novel enhanced applications of QSPR models: Temperature dependence of aqueous solubility. *J. Comput. Chem.* 37 (22): 2045-2051.
85. J. Jaworska, N. Nikolova-Jeliazkova, T. Aldenberg (2005) QSAR applicability domain estimation by projection of the training set descriptor space: a review, *Altern. Lab. Anim.*, 33: 445-459.
86. Bishop CM, M. Svensen M, Williams CKI. (1998). Developments of the Generative Topographic Mapping Neurocomputing 21: 203-224
87. Gaspar H, Sidorov P., Horvath D., Baskin I., Marcou G., Varnek A. (2016) Generative Topographic Mapping Approach to Chemical Space Analysis, *in* *Frontiers in Molecular Design and Chemical Information Science*
88. Kohonen T. (2001). Self-organizing maps. Berlin, New York, Springer.
89. Klimenko K, Marcou G, Horvath, D, Varnek A. (2016). Chemical Space Mapping and Structure-Activity Analysis of the ChEMBL Antiviral Compound Set. *J. Chem. Inf. Model.* 56 (8): 1438-1454.
90. Horvath D, Marcou G, Varnek A (2009). Predicting the predictability: a unified approach to the applicability domain problem of QSAR models. *J. Chem. Inf. Model.* 49 (7): 1762-1776.
91. Huang N, Shoichet BK, Irwin JJ. (2006). Benchmarking sets for molecular docking. *J. Med. Chem.* 49 (23): 6789-6801.
92. Sushko I, Novotarskyi S, Korner R, Pandey AK, Cherkasov A, Li J, Gramatics P, Hansen K, Schroeter T, Muller KR, XI L, Liu H, Yao X, Oberg T, Hormozdiari F, Dao P, Sahinalp C, Todeschini R, Polishchuk P, Artemenko A, Kuz'min V, Martin TM, Young DM, Fourches D, Muratov E, Tropsha A, Baskin I, Horvath D, Marcou G, Muller C, Varnek A, Prokopenko VV, Tetko IV (2010). Applicability domains for classification problems: Benchmarking of distance to models for Ames mutagenicity set. *J. Chem. Inf. Model.* 50 (12): 2094-2111.
93. Czodrowski P. (2014). Count on kappa. *J. Comput. Aided Mol. Des.* 28 (11): 1049-1055
94. Kohavi R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. Proceedings of the 14th international joint conference on Artificial intelligence, Montreal, Quebec, Morgan Kaufmann Publishers Inc

95. Langer T, Wolber G. (2004). Pharmacophore definition and 3D searches. *Drug Discovery Today: Technologies* 1 (3): 203-207.
96. Poduch E, Bello AM, Tang S, Fujihashi M, Pai EF, Kotra LP (2006). Design of inhibitors of orotidine monophosphate decarboxylase using bioisosteric replacement and determination of inhibition kinetics. *J. Med. Chem.* 49 (16): 4937-4945.
97. Leach AR, Gillet VJ, Lewis RA, Taylor R. (2010). Three-Dimensional Pharmacophore Methods in Drug Discovery. *J. Med. Chem.* 53 (2): 539-558.
98. Langer T, Hoffmann RD. (2006). Pharmacophores and pharmacophore searches. *Methods and principles in medicinal chemistry* 32:395
99. Wolber G, Dornhofer AA, Langer T. (2006). Efficient overlay of small organic molecules using 3D pharmacophores. *J. Comput.-Aided Mol. Des.* 20 (12): 773-788.
100. Wolber G, Seidel T, Bendix F, Langer T (2008). Molecule-pharmacophore superpositioning and pattern matching in computational drug design. *Drug Discovery Today* 13 (1-2): 23-29.
101. Wolber G, Seidel T, Bendix F, Ibis G, Dornhofer AA, Biely M, Adaktylos P, Kosara R. (2010). LigandScout, Inte:Ligand GmbH, Vienna, Austria.
102. Mysinger MM, Carchia M, Irwin JJ, Shoichet BK. (2012). Directory of useful decoys, enhanced (DUD-E): better ligands and decoys for better benchmarking. *J. Med. Chem.* 55 (14): 6582-6594.
103. Filimonov DA. (2006). Prediction of biological activity spectra for organic compounds. *Russian Chemical Journal* 50 (2): 66-75.
104. Rognan D. (2005). BioinfoDB : un inventaire de molécules commercialement disponibles à des fins de criblage biologique. *La Gazette du CINES*: 1-4.
105. <https://pubchem.ncbi.nlm.nih.gov/>.
106. <http://zinc.docking.org/>.
107. doi: 10.6019/CHEMBL.database.19.
108. Berthold MR, Cebron N., Dill, F., Gabriel, T. R., Kötter, T., Meinel, T., Ohl, P., Sieb, C., Thiel, K. (2007). *KNIME: The Konstanz Information Miner In Data Analysis, Machine Learning and Applications*, C. Preisach, Burkhardt, H., Schmidt-Thieme, L., Decker, R. Berlin, Springer: 319-326.
109. Witten I, Frank E. (2005). *Data Mining: Practical machine learning tools and techniques*. San Francisco, Morgan Kaufmann.
110. R development Core Team (2007). *R: A language and environment for statistical computing*. <http://www.R-project.org>.
111. Gilbert, D. (2005). *JFreeChart Developer Guide*, Object Renery Limited, Berkeley, California, <http://www.jfree.org/jfreechart>.
112. Varin T, Schuffenhauer A, Ertl P, Renner S. (2011). Mining for bioactive scaffolds with scaffold networks: improved compound set enrichment from primary screening data. *J. Chem. Inf. Model.* 51 (7): 1528-1538.
113. Schuffenhauer A, Ertl P, Roggo S, Wetzels S, Koch MA, Waldmann H (2007). The scaffold tree--visualization of the scaffold universe by hierarchical scaffold classification. *J. Chem. Inf. Model.* 47 (1): 47-58.
114. Yalkowsky, S. H. and Y. He (2003). *Handbook of aqueous solubility data*. Boca Raton, Fla., CRC Press.
115. Nordstrom FL, Rasmuson AC. (2009). Prediction of solubility curves and melting properties of organic and pharmaceutical compounds. *Eur J Pharm Sci* 36 (2-3): 330-344.
116. TableCurve 2D version 5.01, trial version available from: <https://systatsoftware.com/products/tablecurve-2d/>.

117. Varnek A, Gaudin C, Marcou G, Baskin I, Pandey AK, Tetko IV. Inductive transfer of knowledge: application of multi-task learning and feature net approaches to model tissue-air partition coefficients *J. Chem. Inf. Model.* 2009, 49, 133–144.
118. Katritzky AR, Maran U, Lobanov VS, Karelson M. Structurally diverse quantitative structure–property relationship correlations of technologically relevant physical properties, *J. Chem. Inf. Comput. Sci.* 2000, 40, 1.
119. Sun F, Kang H, Baoshu Liu K, Zhang B. Solubility of chlocyphos in different solvents *Fluid Phase Equilibr.* 2012, 330, 12–16.
120. Shi X, Zhou C, Gao Y, Chen X. Measurement and Correlation for Solubility of (S) - (+) - 2,2 - Dimethylcyclopropane Carbox Amide in Different Solvents *Chinese J. Chem. Eng.* 2006, 14, 547–550.
121. Yang W, Wang K, Hu Y, Shen F, Feng J, Solubility of l-tartaric Acid in Ethanol, Propanol, Isopropanol, n-Butanol, Acetone and Acetonitrile *J. Solution. Chem.*, 2013, 42, 485-493.
122. Yang G.-D, Li C, Zeng A.-G, Qu Q.-H, Yang X, Bian X.-L. Solubility of osthole in a binary system of ethanol and water *Fluid Phase Equilibr.*, 2012, 325, 41–44.
123. Yang G.-D, Li C, Zeng A.-G, Guo Y.-L, Yang X, Xing J.-F, Solubility of imperatorinin ethanol plus water mixtures *J. Mol. Liq.* 2012, 167, 86–88
124. Klamt A. Conductor-like Screening Model for Real Solvents: A New Approach to the Quantitative Calculation of Solvation Phenomena *J. Phys. Chem.* 1995, 99, 2224-2235
125. Antonini I, Polucci P, Kelland LR, Menta E, Pescalli N, Martelli S. 2,3-Dihydro-1H,7H-pyrimido [5,6,1-de]acridine-1,3,7-trione Derivatives, a Class of Cytotoxic Agents Active on Multidrug-Resistant Cell Lines: Synthesis, Biological Evaluation, and Structure-Activity Relationships, *J. Med. Chem.*, 42 (1999) 2535 – 2541.
126. www.chemaxon.com/jchem/doc/user/standardizer.html ChemAxon Standardizer. (accessed Feb. 2009)
127. Calculator Plugins were used for structure property prediction and calculation, Marvin 6.1.4, 2013, ChemAxon (<http://www.chemaxon.com>)
128. <https://www.antiviralintelistrat.com/>
129. Hockova D, Holy A, Masojdkova M, Andrei G, Snoeck R, De Clercq E, Balzarini J. (2003). 5-Substituted-2,4-diamino-6-[2 (phosphonomethoxy)ethoxy]pyrimidines-acyclic nucleoside phosphonate analogues with antiviral activity. *J. Med. Chem.* 46 (23): 5064-5073.
130. Mugnaini C, Manetti F, Este JA, Clotet-Codina I, Maga G, Cancio R, Botta M, Corelli F. (2006). Synthesis and biological investigation of S-aryl-S-DABO derivatives as HIV-1 inhibitors. *Bioorg. Med. Chem. Lett.* 16 (13): 3541-3544.
131. Kumar R, Nath M, Tyrrell DL. Design and Synthesis of Novel 5-Substituted Acyclic Pyrimidine Nucleosides as Potent and Selective Inhibitors of Hepatitis B Virus. *J. Med. Chem.* 2002, 45, 2032-2040.
132. Lin TS, Zhu JL, Dutschman GE, Cheng YC, Prusoff WH. Syntheses and Biological Evaluations of 3'-Deoxy-3'-C-Branched-Chain-Substituted Nucleosides. *J. Med. Chem.* 1993, 36, 353-362.
133. DeGoey DA, Randolph JT, Liu D, Pratt J, Hutchins C, Donner P, Krueger AC, Matulenko M, Patel S, Motter CE, Nelson L, Keddy R, Tufano M, Caspi D, Krishnan P, Mistry N, Koev G, Reisch TJ, Mondal R, Pilot-Matias T, Gao Y, Beno DW, Maring CJ, Molla A, Dumas E, Campbell A, Williams L, Collins C, Wagner R, Kati WM. Discovery of ABT-267, a Pan-Genotypic Inhibitor of HCV NS5A. *J. Med. Chem.*, 2014 57, 2047-2057.

134. Govorkova EA, Ilyushina NA, Boltz DA, Douglas A, Yilmaz N, Webster RG. (2007). Efficacy of oseltamivir therapy in ferrets inoculated with different clades of H5N1 influenza virus. *Antimicrob. Agents Chemother.* 51 (4): 1414-1424.
135. Tonelli M, Boido V, Canu C, Sparatore A, Sparatore F, Paneni MS, Fermeglia M, Pricl S, La Colla P, Casula L, Ibba C, Collu D, Loddo R. Antimicrobial and cytotoxic arylazoenamines. Part III: antiviral activity of selected classes of arylazoenamines. *Bioorg. Med. Chem.* 2008, 16, 8447-8465.
136. Giliberti G, Ibba C, Marongiu E, Loddo R, Tonelli M, Boido V, Laurini E, Posocco P, Fermeglia M, Pricl S. Synergistic Experimental/Computational Studies on Arylazoenamine Derivatives that Target the Bovine Viral Diarrhea Virus RNA-Dependent RNA Polymerase. *Bioorg. Med. Chem.* 2010, 18, 6055-6068.
137. (2012). Virus Taxonomy: Ninth Report of the International Committee on Taxonomy of Viruses.: 1327.
138. Whitley RJ, Roizman B. Herpes Simplex Virus Infections. *Lancet*, 2001, 357, 1513-1518.
139. Weininger D. SMILES, A Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules. *J. Chem. Inf. Comput. Sci.* 1988, 28, 31-36.
140. Gaspar HA, Baskin II, Marcou G, Horvath D, Varnek A. Chemical Data Visualization and Analysis with Incremental Generative Topographic Mapping: Big Data Challenge. *J. Chem. Inf. Model.* 2014, 55, 84-94
141. Gaspar HA, Baskin II, Marcou G, Horvath D, Varnek A. GTM-Based QSAR Models and Their Applicability Domains. *Mol. Inf.* 2015, 34, 348-356
142. Sidorov P, Gaspar H, Marcou G, Varnek A, Horvath D. Mappability of Drug-Like Space: Towards a Polypharmacologically Competent Map of Drug-Relevant Compounds. *J. Comput.-Aided Mol. Des.* 2015, 29, 1087-1108.
143. Ruggiu F, Marcou G, Varnek A, Horvath D. Isida Property-labelled Fragment Descriptors. *Mol. Inf.* 2010, 29, 855-868.
144. Buckheit RW, Buckheit KW, Sturdevant CB, Buckheit RW. Selection and Characterization of Viruses Resistant to The Dual Acting Pyrimidinedione Entry and Non-Nucleoside Reverse Transcriptase Inhibitor IQP-0410. *Antiviral Res.* 2013, 100, 382-391.
145. Yu M, Corsa AC, Xu SM, Peng B, Gong RY, Lee YJ, Chan KT, Mo HM, Delaney W, Cheng GF. In Vitro Efficacy of Approved and Experimental Antivirals Against Novel Genotype 3 Hepatitis C Virus Subgenomic Replicons. *Antiviral Res.* 2013, 100, 439-445.
146. Peng HK, Chen WC, Lin YT, Tseng CK, Yang SY, Tzeng CC, Lee JC, Yang S. C. Anti-hepatitis C Virus RdRp Activity and Replication of Novel Anilinobenzothiazole Derivatives. *Antiviral Res.* 2013, 100, 269-275.
147. Biswas S, Swift M, Field HJ. High Frequency of Spontaneous Helicase-Primase Inhibitor (BAY 57-1293) Drug-Resistant Variants in Certain Laboratory Isolates Of HSV-1. *Antivir. Chem. Chemother.* 2007, 18, 13-23
148. Ivachtchenko AV, Ivanenkov YA, Mitkin OD, Yamanushkin PM, Bichko VV, Leneva IA, Borisova OV A Novel Influenza Virus Neuraminidase Inhibitor AV5027. *Antiviral Res.* 100, 698-708.
149. Kim M, Kim SY, Lee HW, Shin JS, Kim P, Jung YS, Jeong HS, Hyun JK, Lee CK. Inhibition of Influenza Virus Internalization by (-)-epigallocatechin-3-gallate. *Antiviral Res.* 2013, 100, 460-572.
150. Martinez-Gil L, Alamares-Sapuay JG, Ramana Reddy MV, Goff PH, Premkumar Reddy E., Palese PA. Small Molecule Multi-Kinase Inhibitor Reduces Influenza A Virus Replication by Restricting Viral RNA Synthesis. *Antiviral Res.* 2013, 100, 29-37.

151. Haasbach E, Reiling SJ, Ehrhardt C, Droebner K, Ruckle A, Hrinčius ER, Leban J, Strobl S, Vitt D, Ludwig S, Planz O The NF-kappaB Inhibitor SC75741 Protects Mice Against Highly Pathogenic Avian Influenza A Virus. *Antiviral Res.* 2013, 99, 336-344.
152. Evans BE, Rittle KE, Bock MG, DiPardo RM, Freidinger RM, Whitter WL, Lundell GF, Veber DF, Anderson PS Methods for Drug Discovery: Development of Potent, Selective, Orally Effective Cholecystokinin Antagonists. *J. Med. Chem.* 1988, 31, 2235-2246.
153. Kubinyi H. Privileged Structures and Analogue-Based Drug Discovery. In *Analogue-based Drug Discovery*, Fischer JGR., Ed. Wiley-VCH Verlag-GmbH & Co. KGaA, Weinheim, Germany, 2006, pp 53-68.
154. Shi J, Zhou L, Amblard F, Bobeck DR, Zhang H, Liu P, Bondada L, McBrayer TR, Tharnish PM, Whitaker T, Coats SJ, Schinazi RF. Synthesis and Biological Evaluation of New Potent and Selective HCV NS5A Inhibitors. *Bioorg. Med. Chem. Lett.* 2012, 22, 3488-3491
155. Abdel-Magid AF. Halting HCV Replication with NS5A Inhibitors and NS5B Polymerase Inhibitors: Effective New Treatments of HCV Infection. *ACS Med. Chem. Lett.*, 2013, 5, 234-237
156. White LK, Yoon JJ, Lee JK, Sun AM, Du YH, Fu H. Snyder, J. P., Plemper, R. K. Nonnucleoside Inhibitor of Measles Virus RNA-Dependent RNA Polymerase Complex Activity. *J. Antimicrob. Agents Chemother* 2007, 51, 2293-2303.
157. Yu W, Goddard C, Clearfield E, Mills C, Xiao T, Guo H, Morrey JD, Motter NE, Zhao K, Block TM, Cuconati A, Xiaodong Xu Synthesis, and Biological Evaluation of Triazolo-pyrimidine Derivatives as Novel Inhibitors of Hepatitis B Virus Surface Antigen (HBsAg) Secretion. *J. Med. Chem.*, 2011, 54 (16), pp 5660–5670
158. Prichard MN, Hartline CB, Harden EA, Daily SL, Beadle JR, Valiaeva N, Kern ER, Hostetler KY. Inhibition of Herpesvirus Replication by Hexadecyloxypropyl Esters of Purine- and Pyrimidine-Based Phosphonomethoxyethyl Nucleoside Phosphonates. *Antimicrob. Agents Chemother.* 2008, 52, 4326-4330
159. Baszczyński O, Jansa P, Dracinsky M, Klepetarova B, Holy A, Votruba I, De Clercq E, Balzarini J, Janeba Z. Synthesis and Antiviral Activity of N9- [3-fluoro-2-(phosphonomethoxy)propyl] Analogues Derived from N6-substituted Adenines and 2,6-Diaminopurines. *Bioorg. Med. Chem.* 2011, 19, 2114-2124.
160. Zhou SM, Drach JC, Prichard MN, Zemlicka J. (Z)- and (E)-2- (1,2-Dihydroxyethyl)methylenecyclopropane Analogues of 2'-Deoxyadenosine and 2'-Deoxyguanosine. Synthesis of All Stereoisomers, Absolute Configuration, and Antiviral Activity. *J. Med. Chem.* 2009, 52, 3397-3407.
161. Diez-Torrubia A, Cabrera S, De Castro S, Garcia-Aparicio C, Mulder G, De Meester I, Camarasa MJ, Balzarini J, Velazquez S. Novel Water-Soluble Prodrugs of Acyclovir Cleavable by the Dipeptidyl-Peptidase IV (DPP IV/CD26) Enzyme. *Eur. J. Med. Chem.* 2013, 70, 456-468.
162. Gudmundsson KS, Johns BA. Imidazo [1,2-a]pyridines with Potent Activity Against Herpesviruses. *Bioorg. Med. Chem. Lett.* 2007, 17, 2735-2739.
163. Prado-Prado FJ, De la Vega OM, Uriarte E, Ubeira FM, Chou KC, Gonzalez-Diaz H. Unified QSAR Approach to Antimicrobials. 4. Multi-target QSAR Modeling and Comparative Multi-Distance Study of the Giant Components of Antiviral Drug-Drug Complex Networks. *Bioorg. Med. Chem.* 2009, 17, 569-575.
164. Krecmerova M, Holy A, Piskala A, Masojdkova M, Andrei G, Naesens L, Neyts J, Balzarini J, De Clercq E, Snoeck R. Antiviral Activity of Triazine Analogues

- of 1- (S)- [3-hydroxy-2- (phosphonomethoxy)propyl]cytosine (Cidofovir) and Related Compounds. *J. Med. Chem.* 2007, 50, 1069-1077
165. Wang LJ, Geng CA, Ma YB, Huang XY, Luo J, Chen H, Guo RH, Zhang X M, Chen JJ. Synthesis, Structure-Activity Relationships and Biological Evaluation of Caudatin Derivatives as Novel Anti-Hepatitis B Virus Agents. *Bioorg. Med. Chem.* 2012, 20, 2877-2888.
166. Wang ZQ, Bennett EM, Wilson DJ, Salomon C, Vince R. Rationally Designed Dual Inhibitors of HIV Reverse Transcriptase and Integrase. *J. Med. Chem.* 2007, 50, 3416-3419.
167. Yang J, Zhang F, Li JR, Chen G, Wu SW, Ouyang WJ, Pan W, Yu R, Yang JX, Tien P. Synthesis and Antiviral Activities of Novel Gossypol Derivatives. *Bioorg. Med. Chem. Lett.* 2012, 22, 1415-1420.
168. Crosby IT, Bourke DG, Jones ED, Jaynes TP, Cox S, Coates JAV, Robertson AD Antiviral agents 3. Discovery of a novel small molecule non-nucleoside inhibitors of Hepatitis B virus (HBV). *Bioorg. Med. Chem. Lett.* 2011, 21, 6, 1644-1648.
169. <https://sourceforge.net/projects/scaffoldhunter/>
170. Chezal JM, Paeshuysse J, Gaumet V, Canitrot D, Maisonia A, Lartigue C, Gueiffier A, Moreau E, Teulade JC, Chavignon O, Neyts J. Synthesis and Antiviral Activity of an Imidazo [1,2-a]pyrrolo [2,3-c]pyridine Series Against The Bovine Viral Diarrhea Virus. *Eur. J. Med. Chem.* 2010, 45, 2044-2047.
171. Ciustea M, Silverman JE, Druck Shudofsky A. M., Ricciardi, R. P. (2008). Identification of non-nucleoside DNA synthesis inhibitors of vaccinia virus by high-throughput screening. *J. Med. Chem.* 51 (20): 6563-6570.
172. Wilhelmsson LM, Kingi N, Bergman J. (2008). Interactions of antiviral indolo [2,3-b]quinoxaline derivatives with DNA. *J. Med. Chem.* 51 (24): 7744-7750
173. Upadhyay KMA, Loddio R, La Colla P, Virsodiya V, Trivedi J, Chaniyara R. (2013). Syntheses and in vitro biological screening of 1-aryl-10H- [1,2,4]triazolo [3',4':3,4] [1,2,4]triazino [5,6-b]indoles. *Med. Chem. Res.* 22 (8): 3675–3686.
174. Denny WA, Atwell GJ, Baguley BC, Wakelin LP. (1985). Potential antitumor agents. Synthesis and antitumor activity of new classes of diacridines: importance of linker chain rigidity for DNA binding kinetics and biological activity. *J. Med. Chem.* 28 (11): 1568-1574.
175. Chang J.-W. A New Cell-Based Clustering Method for High-Dimensional Data Mining Applications. In Knowledge-Based Intelligent Information and Engineering Systems: Part I of Proceedings of 9th International Conference, KES 2005, Melbourne, Australia, September 14-16, 2005,
176. Khosla R, Howlett RJ, Jain LC. Eds. Springer Berlin Heidelberg: Berlin, Heidelberg, Germany, 2005, pp 391-397.
177. Hu Y, Stumpfe D, Bajorath J. Lessons Learned from Molecular Scaffold Analysis *J. Chem. Inf. Model.* 2011, 51, 1742–1753

APPENDICES

APPENDIX A Supplementary material to QSPR modeling of aqueous solubility	126
Table S1 Regression analysis data.....	126
Table S2 Solubility coefficient training set data.....	129
Table S3 External tests compounds prediction result.....	133
APPENDIX B Supplementary material to computer-aided design of new antiviral compounds.....	137
Table S4 Training set compounds – antiviral intercalators.....	137
Table S5 All results of cytotoxicity and antiviral activity measured by Flow cytometry analysis.....	159
Figure S1 Cells viability within 110 h in the presence of 10, 24, 15, 19, 17 at different concentration.....	161
APPENDIX C Published and submitted articles.....	163

APPENDIX A Supplementary material to QSPR modeling of aqueous solubility

Table S1 Regression analysis data

CAS	F-test	equation
57-50-1	303.4	$lgSW = c + k \sqrt{T}$
	521.8	$lgSW = c + k^2 \sqrt{T^3}$
	549.1	$lgSW = c + k \sqrt{T} \ln T$
	891.8	$lgSW = c + kT / \ln T$
	1098.0	$lgSW = c + kT \ln T$
	1272.1	$lgSW = c + kT$
100-09-4	320.1	$lgSW = c + k \sqrt{T} \ln T$
	320.7	$lgSW = c + kT^2$
	570.5	$lgSW = c + kT / \ln T$
	1219.9	$lgSW = c + kT$
	1831.8	$lgSW = c + k^2 \sqrt{T^3}$
	3914.7	$lgSW = c + kT \ln T$
120-12-7	262.8	$lgSW = c + k^2 \sqrt{T^5}$
	266.0	$lgSW = c + k / \ln T$
	269.5	$lgSW = c + k / \sqrt{T}$
	375.6	$lgSW = c + kT^2 \ln T$
	528.8	$lgSW = c + kT^2$
	595.3	$lgSW = c + k \ln T$
	1020.0	$lgSW = c + k(\ln T)^2$
	1504.6	$lgSW = c + k^2 \sqrt{T^3}$
	1948.4	$lgSW = c + k \sqrt{T}$
	3210.1	$lgSW = c + k \sqrt{T} \ln T$
	3499.6	$lgSW = c + kT \ln T$
	4871.3	$lgSW = c + kT / \ln T$
	5320.5	$lgSW = c + kT$
	1202-25-1	277.1
546.6		$lgSW = c + k(\ln T)^2$
790.2		$lgSW = c + kT$
2454.1		$lgSW = c + k \sqrt{T}$
3456.7		$lgSW = c + kT / \ln T$
387745.9		$lgSW = c + k \sqrt{T} \ln T$
141-82-2	325.6	$lgSW = c + kT \ln T$
	329.4	$lgSW = c + k(\ln T)^2$
	690.2	$lgSW = c + kT$
	748.1	$lgSW = c + k \sqrt{T}$
	1224.0	$lgSW = c + kT / \ln T$
	1328.5	$lgSW = c + k \sqrt{T} \ln T$
461-58-5	396.2	$lgSW = c + kT \ln T$
	442.1	$lgSW = c + k(\ln T)^2$
	1065.1	$lgSW = c + kT$

	1613.8	$lgSW = c + k \sqrt{T}$
	4008.6	$lgSW = c + kT/\ln T$
	11023.1	$lgSW = c + k \sqrt{T} \ln T$
591-27-5	331.8	$lgSW = c + kT$
	385.5	$lgSW = c + kT^2 \ln T$
	646.7	$lgSW = c + kT \ln T$
	662.1	$lgSW = c + kT^2$
	1356.5	$lgSW = c + k^2 \sqrt{T^3}$
2051-24-3	270.1	$lgSW = c + kT^2$
	276.5	$lgSW = c + kT/\ln T$
	471.5	$lgSW = c + kT$
	1057.9	$lgSW = c + kT \ln T$
	1809.0	$lgSW = c + k^2 \sqrt{T^3}$
68-96-2	277.4	$lgSW = c + kT$
	371.1	$lgSW = c + k \ln T$
	393.3	$lgSW = c + kT/\ln T$
	524.4	$lgSW = c + k \sqrt{T} \ln T$
	629.8	$lgSW = c + k(\ln T)^2$
	651.3	$lgSW = c + k \sqrt{T}$
108-90-7	273.2	$lgSW = c + k \sqrt{T}$
	392.6	$lgSW = c + k \sqrt{T} \ln T$
	471.1	$lgSW = c + k^2 \sqrt{T^3}$
	489.2	$lgSW = c + kT/\ln T$
	581.5	$lgSW = c + kT$
	582.4	$lgSW = c + kT \ln T$
88-72-2	629.7	$lgSW = c + kT^2$
	936.3	$lgSW = c + kT$
	4551.2	$lgSW = c + k \sqrt{T} \ln T$
	5716.5	$lgSW = c + kT/\ln T$
562-49-2	305.3	$lgSW = c + kT \ln T$
	581.4	$lgSW = c + kT^2 \ln T$
	1305.7	$lgSW = c + k^2 \sqrt{T^3}$
	2383.9	$lgSW = c + kT^2$
554-12-1	285.1	$lgSW = c + k(\ln T)^2$
	309.6	$lgSW = c + kT/\ln T$
	367.7	$lgSW = c + k \sqrt{T}$
	376.4	$lgSW = c + k \sqrt{T} \ln T$
584-02-1	253.0	$lgSW = c + kT \ln T$
	448.0	$lgSW = c + kT$
	600.2	$lgSW = c + k \ln T$
	743.9	$lgSW = c + kT/\ln T$
	767.9	$lgSW = c + k \sqrt{T} \ln T$
	1303.3	$lgSW = c + k(\ln T)^2$
	2303.8	$lgSW = c + k \sqrt{T}$
98-01-1	302.1	$lgSW = c + kT/\ln T$
	307.8	$lgSW = c + kT^2 \ln T$

	462.8	$lgSW = c + kT$
	470.9	$lgSW = c + kT^2$
	795.6	$lgSW = c + kT \ln T$
	1073.8	$lgSW = c + k\sqrt[2]{T^3}$
111-14-8	292.0	$lgSW = c + kT$
	393.3	$lgSW = c + k \ln T$
	612.5	$lgSW = c + kT / \ln T$
	1704.6	$lgSW = c + k \sqrt{T} \ln T$
	2690.9	$lgSW = c + k(\ln T)^2$
	45035.9	$lgSW = c + k \sqrt{T}$
110-15-6	317.4	$lgSW = c + k\sqrt[2]{T^3}$
	454.3	$lgSW = c + k \sqrt{T} \ln T$
	805.0	$lgSW = c + kT / \ln T$
	807.3	$lgSW = c + kT \ln T$
	1214.8	$lgSW = c + kT$
302-72-7	450.5	$lgSW = c + kT \ln T$
	626.8	$lgSW = c + k \sqrt{T}$
	1322.0	$lgSW = c + kT$
	1723.4	$lgSW = c + k \sqrt{T} \ln T$
	2898.3	$lgSW = c + kT / \ln T$

Table S2. Solubility coefficient training set data

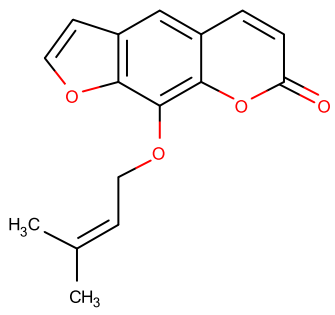
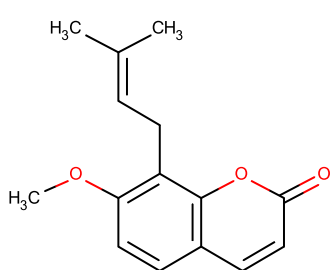
CAS number	Observed k	Predicted k
100-00-5	0.233	0.227
100-02-7	0.275	0.244
100-09-4	0.26	0.262
103-84-4	0.241	0.231
106-46-7	0.214	0.23
106-89-8	0.139	0.11
106-93-4	0.156	0.181
107-13-1	0.132	0.148
107-35-7	0.212	0.171
107-87-9	-0.174	-0.03
108-10-1	-0.151	0.006
108-46-3	0.181	0.23
108-78-1	0.251	0.224
108-86-1	0.168	0.217
108-90-7	0.22	0.178
108-95-2	0.192	0.191
109-94-4	0.141	0.039
110-15-6	0.247	0.196
110-16-7	0.165	0.236
110-17-8	0.253	0.178
110-54-3	-0.182	0.014
110-74-7	-0.129	-0.022
110-82-7	0.097	0.029
110-94-1	0.178	0.202
111-14-8	0.159	0.139
112-38-9	0.189	0.179
115-77-5	0.226	0.147
1185-33-7	-0.183	-0.013
118-92-3	0.255	0.251
118-96-7	0.241	0.185
120-12-7	0.292	0.285
1202-25-1	0.245	0.194
120-80-9	0.21	0.204
120-83-2	0.354	0.28
121-57-3	0.218	0.266
123-30-8	0.22	0.228
123-31-9	0.242	0.227
123-51-3	-0.171	-0.154
123-56-8	0.213	0.193
124-04-9	0.284	0.19
124-07-2	0.168	0.15
129-00-0	0.288	0.293
133-37-9	0.21	0.143
137-32-6	-0.171	-0.127

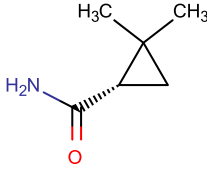
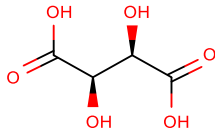
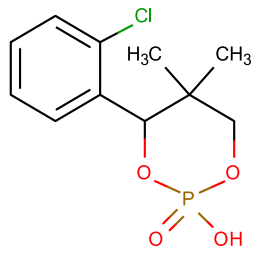
141-78-6	-0.156	-0.031
141-82-2	0.132	0.178
142-62-1	0.128	0.064
144-62-7	0.236	0.167
147-71-7	0.119	0.211
2051-24-3	0.322	0.24
206-44-0	0.288	0.287
217-59-4	0.286	0.286
218-01-9	0.285	0.301
2361-96-8	0.234	0.099
302-72-7	0.168	0.173
30746-58-8	0.244	0.278
315-30-0	0.271	0.215
32598-13-3	0.315	0.256
334-48-5	0.175	0.167
37680-73-2	0.275	0.273
434-03-7	0.24	0.196
461-58-5	0.255	0.175
479-45-8	0.258	0.222
492-62-6	0.109	0.179
50-06-6	0.249	0.226
50-28-2	0.238	0.157
505-48-6	0.292	0.216
50-70-4	0.126	0.187
50-99-7	0.163	0.152
51-28-5	0.256	0.252
51-66-1	0.25	0.233
541-73-1	0.18	0.232
55-21-0	0.249	0.244
553-90-2	0.228	0.172
554-12-1	-0.135	0.008
554-84-7	0.257	0.255
55-63-0	0.188	0.17
56-23-5	0.114	0.185
562-49-2	0.175	-0.007
563-80-4	-0.148	-0.056
56-40-6	0.174	0.187
56-55-3	0.315	0.286
56-84-8	0.248	0.183
57-13-6	0.155	0.204
57-44-3	0.229	0.144
57-50-1	0.106	0.142
57-83-0	0.184	0.166
579-75-9	0.275	0.253
58-22-0	0.242	0.162
584-02-1	-0.184	-0.145

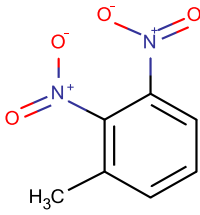
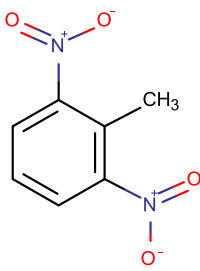
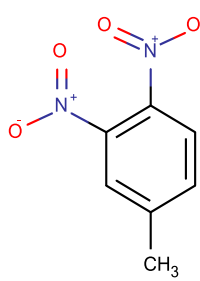
591-27-5	0.274	0.2
59-67-6	0.213	0.229
60-18-4	0.242	0.245
60-29-7	-0.187	-0.085
6032-29-7	-0.188	-0.156
60-35-5	0.144	0.15
613-12-7	0.303	0.283
617-65-2	0.257	0.235
623-37-0	-0.146	-0.063
62-44-2	0.244	0.09
62-53-3	0.139	0.204
62-56-6	0.233	0.17
628-41-1	0.114	0.098
63-74-1	0.295	0.251
65-45-2	0.271	0.25
65-85-0	0.257	0.239
67-66-3	-0.102	0.184
6893-26-1	0.25	0.236
68-96-2	0.249	0.193
6915-15-7	0.143	0.196
69-72-7	0.257	0.264
69-79-4	0.159	0.141
69-93-2	0.258	0.196
71-41-0	-0.173	-0.109
72-14-0	0.261	0.257
73-24-5	0.262	0.243
75-85-4	-0.196	-0.09
76-57-3	0.185	0.161
77-92-9	0.13	0.183
79-20-9	0.109	0.083
832-69-9	0.29	0.292
83-32-9	0.274	0.236
85-01-8	0.286	0.288
86-73-7	0.281	0.276
87-78-5	0.21	0.129
88-72-2	0.175	0.235
88-73-3	0.251	0.238
88-75-5	0.275	0.24
88-89-1	0.209	0.256
88-99-3	0.258	0.257
91-20-3	0.259	0.251
92-52-4	0.275	0.277
94-09-7	0.257	0.193
95-50-1	0.183	0.233
95-55-6	0.138	0.242
98-01-1	0.156	0.188

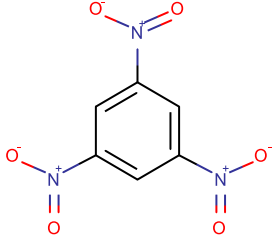
98-18-0	0.275	0.265
98-95-3	0.174	0.221
99-06-9	0.265	0.277
99-96-7	0.284	0.258
99-99-0	0.231	0.197

Table S3 External tests compounds prediction result

Structure,CAS Number and IUPAC Name	T, °C	obs.	our model	COSMO
	15	-4.39	-3.43	-
	20	-4.24	-3.41	-
	25	-4.23	-3.39	-
	30	-4.24	-3.38	-
	35	-4.12	-3.36	-
	40	-4.09	-3.36	-
482-44-0 (9-(3-methylbut-2-en-1-yloxy)furo[3,2-g]chromen-7-one)	45	-4.08	-3.36	-
	50	-4.07	-3.35	-
	55	-4.00	-3.35	-
	15	-4.61	-3.48	-
	20	-4.57	-3.47	-
	25	-4.52	-3.46	-
	30	-4.44	-3.44	-
	35	-4.40	-3.43	-
	40	-4.32	-3.43	-
484-12-8 (7-methoxy-8-(3-methylbut-2-en-1-yl)chromen-2-one)	45	-4.28	-3.43	-
	50	-4.20	-3.43	-
	55	-4.13	-3.43	-

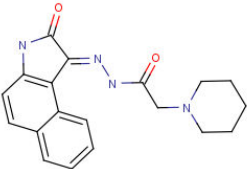
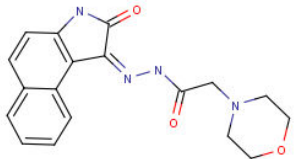
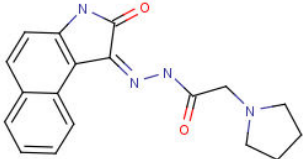
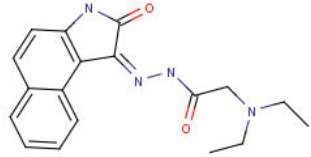

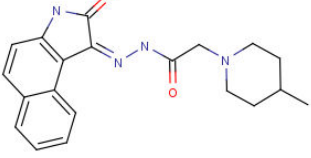


 <p>75885-58-4 ((S)-(+)-2,2-dimethylcyclopropane-1-carboxamide)</p>	22	-0.68	-0.36	-
	35	-0.68	-0.34	-
	42	-0.49	-0.33	-
	54	-0.40	-0.32	-
	59	-0.29	-0.31	-
	63	-0.19	-0.31	-
	81	0.17	-0.29	-
 <p>87-69-4 ((2R,3R)-2,3-dihydroxybutanedioic acid)</p>	15	-0.60	0.31	-
	25	-0.54	0.42	-
	35	-0.44	0.47	-
	45	-0.38	0.47	-
	55	-0.33	0.47	-
 <p>98634-28-7 (4-(2-chlorophenyl)-2-hydroxy-5,5-dimethyl-1,3,2-dioxaphosphorinane 2-oxide)</p>	5	-2.16	-2.90	-
	10	-2.17	-2.90	-
	15	-2.12	-2.90	-
	20	-2.06	-2.89	-
	25	-2.05	-2.87	-
	30	-2.04	-2.87	-
	40	-2.03	-2.85	-

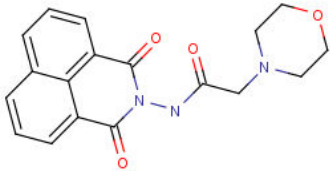

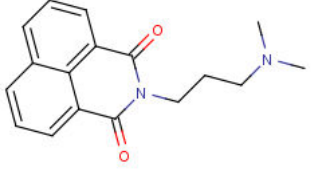

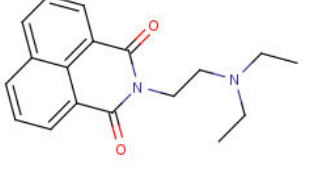
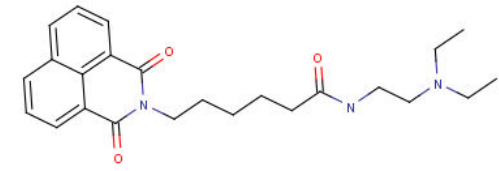
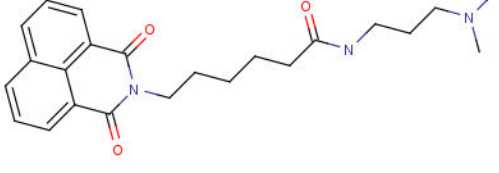
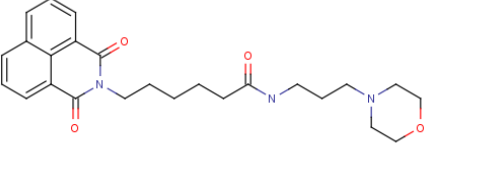
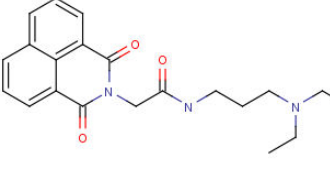
	45	-2.02	-2.84	-
	5	-3.43	-2.9	-3.63
	7	-3.36	-2.9	-3.57
	19	-3.17	-2.87	-3.34
	30	-2.96	-2.82	-3.2
	602-01-7 (1-methyl-2,3-dinitrobenzene)	41	-2.76	-2.79
	5	-3.39	-2.74	-3.58
	7	-3.31	-2.74	-3.54
	19	-3.09	-2.72	-3.25
	30	-2.84	-2.67	-2.99
	606-20-2 (2-methyl-1,3-dinitrobenzene)	41	-2.63	-2.64
	5	-3.42	-2.9	-3.54
	20	-3.13	-2.87	-3.24
	31	-2.95	-2.81	-3.01
	40	-2.75	-2.79	-2.83
610-39-9 (4-methyl-1,2-dinitrobenzene)				

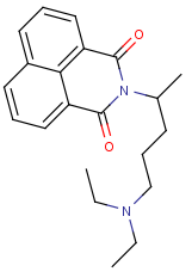
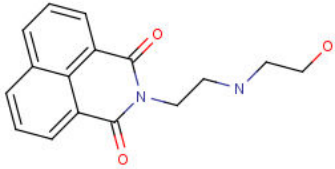
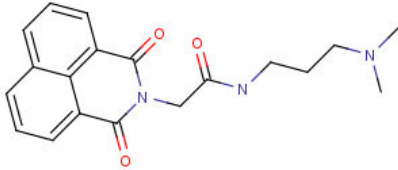
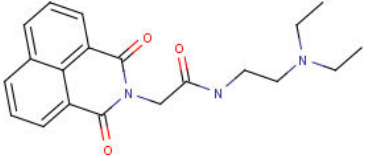
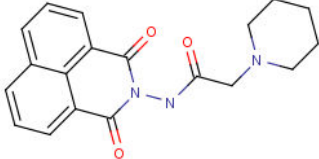
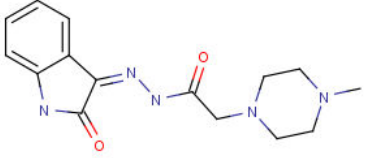
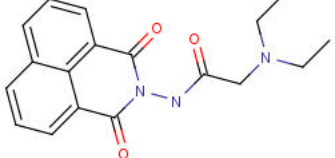
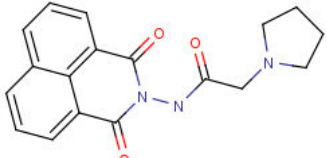
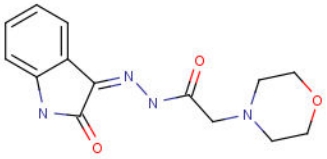
	5	-2.9	-2.7	-4.21
	19	-2.74	-2.68	-3.9
	30	-2.58	-2.59	-3.66
	41	-2.45	-2.46	-3.41
99-35-4 (1,3,5-trinitrobenzene)				
RMSE(total)			0.67	
RMSE(comparison)			0.34	0.57

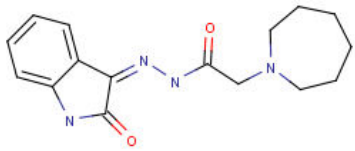
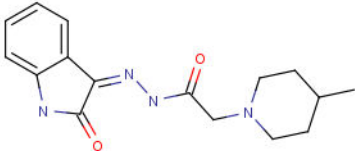
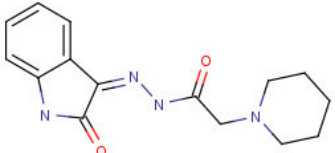
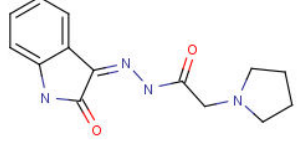
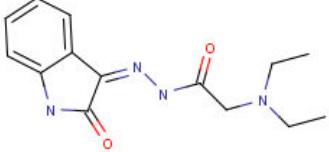
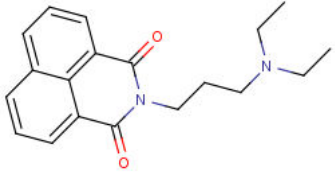
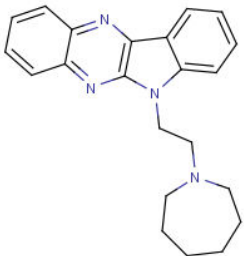
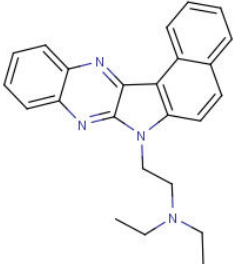
APPENDIX B Supplementary material to computer-aided design of new antiviral compounds

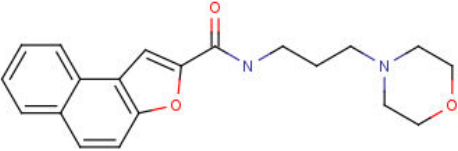
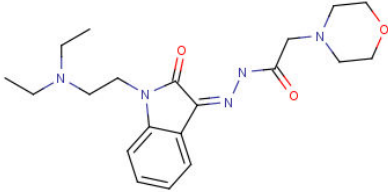
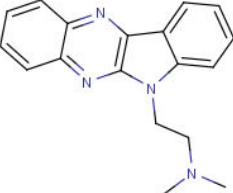
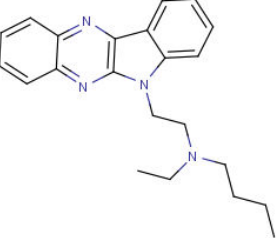
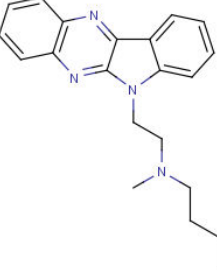
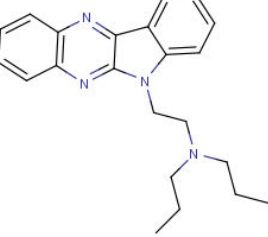
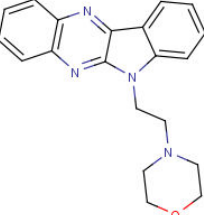
Table S4 Training set compounds – antiviral intercalators

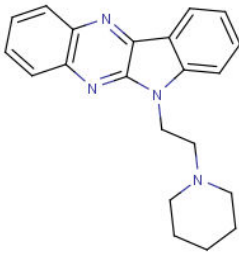
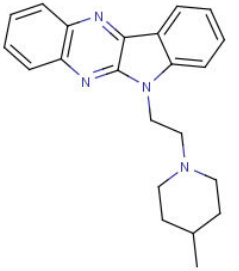
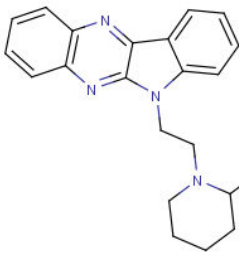
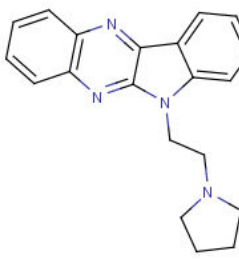
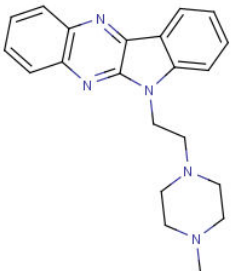
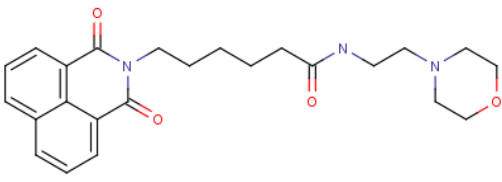
Structure	Cmpd #	Affinity constant(K _a)	Antiviral activity(%)
	1	6.21	10
	2	6.76	15
	3	6.44	10
	4	6.3	30
	5	5.99	70
	6	7.11	40
	7	6.52	20
	8	5.73	

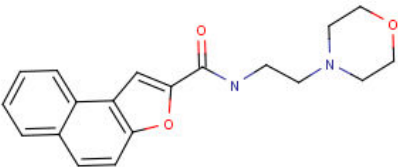
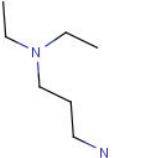
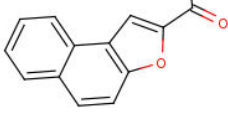
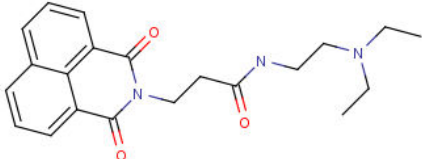
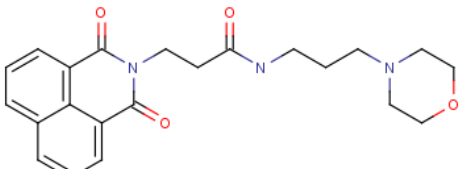
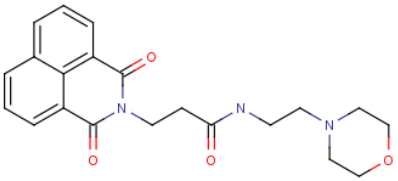
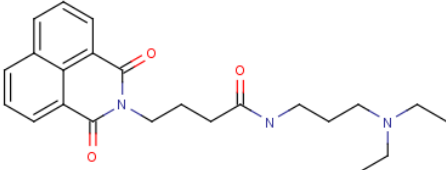
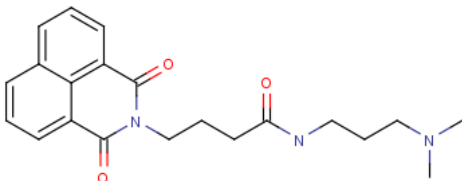
	9	5.34	
	10	5.53	
	11	7.07	
	12	5.43	
	13	6.98	
	14	5.49	
	15	6.1	100
	16	6.03	50
	17	5.31	

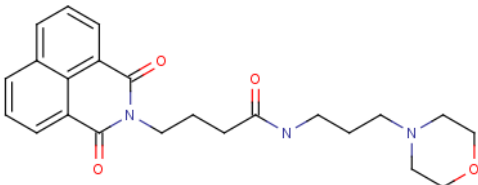
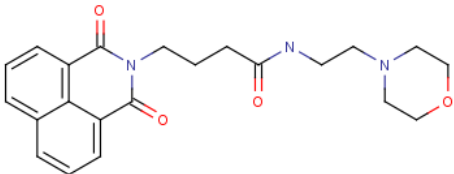
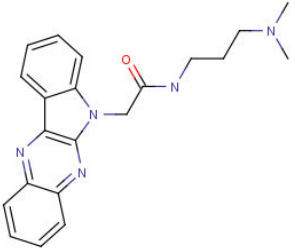
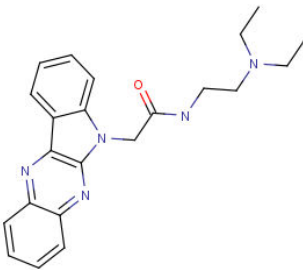
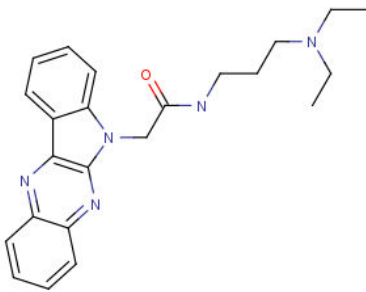
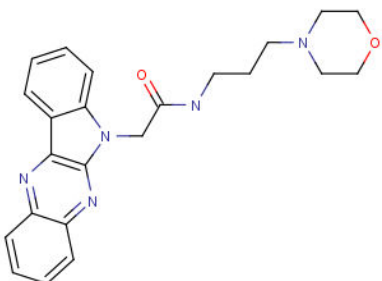
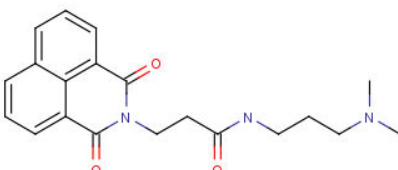
	18	5.5	
	19	5.57	
	20	5.63	
	21	5.54	
	22	6.2	
	23	5.43	80
	24	5.6	
	25	5.41	
	26	5.28	

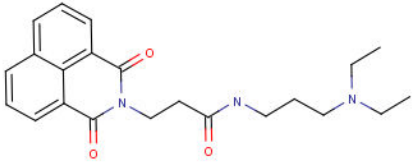
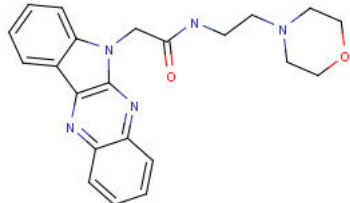
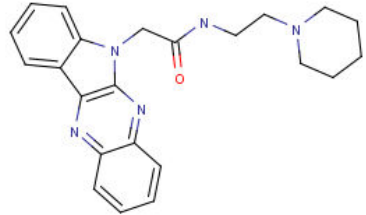
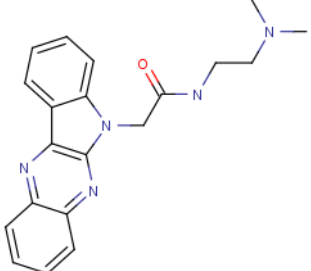
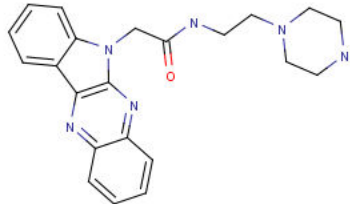
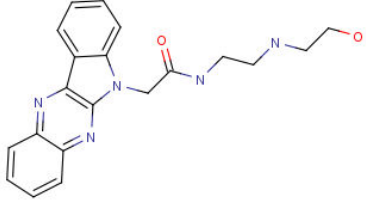
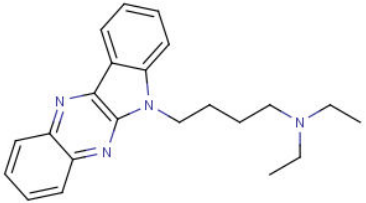
	27	5.49	80
	28	5.4	60
	29	5.64	100
	30	5.58	
	31	5.96	80
	32	6.57	
	33	6.16	90
	34	6.79	14

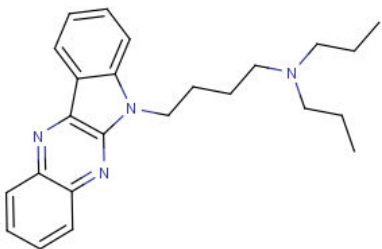
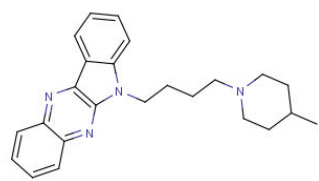
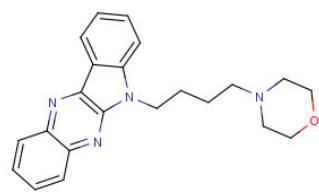
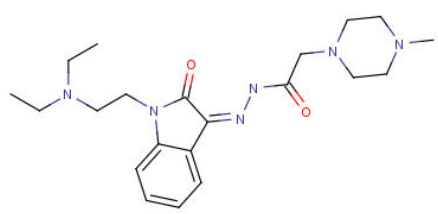
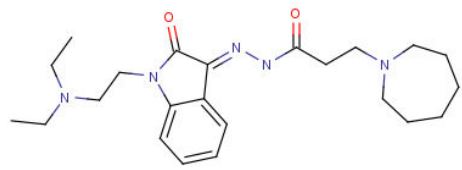
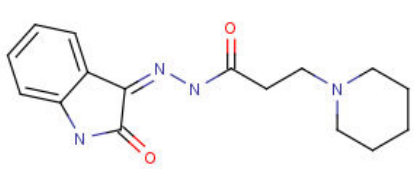
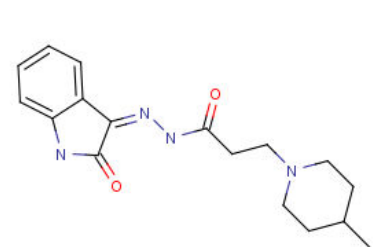
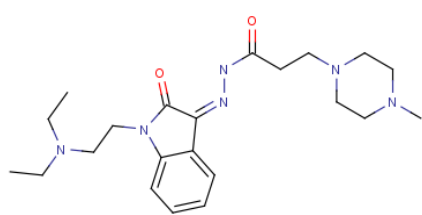
	44	5.64	
	45	6.21	8
	46	5.93	75
	47	6.01	50
	48	5.93	75
	49	6.09	90
	50	5.89	85


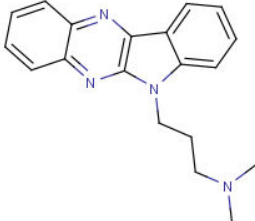
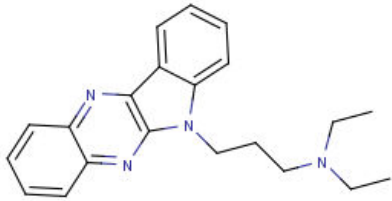
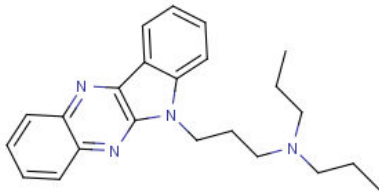
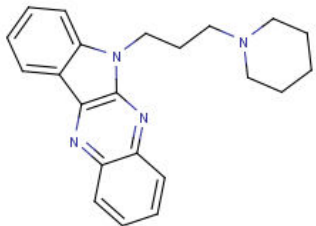
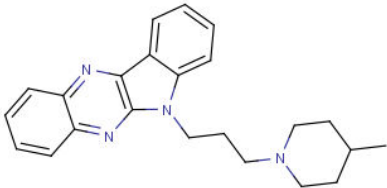
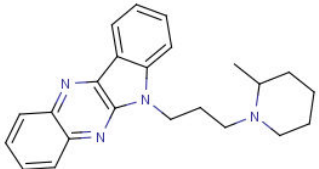
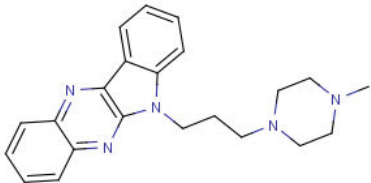
	51	6.01	90
	52	5.98	90
	53	5.19	90
	54	6.05	85
	55	5.87	85
	56	5.53	

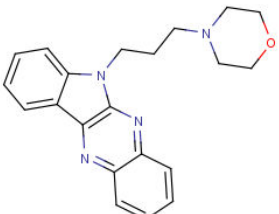
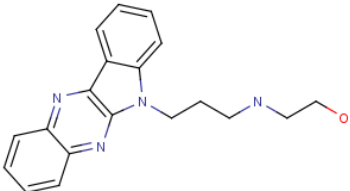
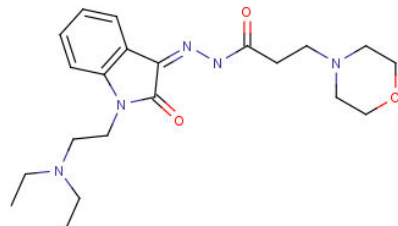
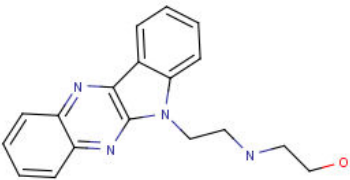
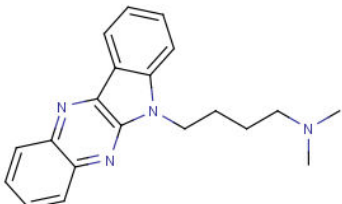

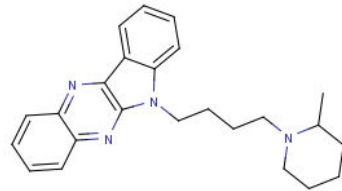
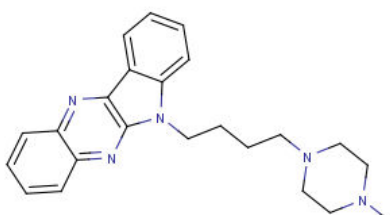
	57	4.58	
	58	6.12	
	59	5.85	20
	60	5.72	30
	61	5.78	5
	62	6.21	15
	63	5.95	
	64	6.13	20

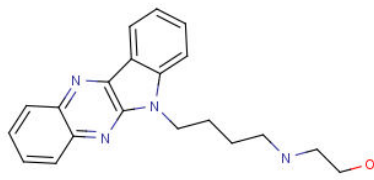
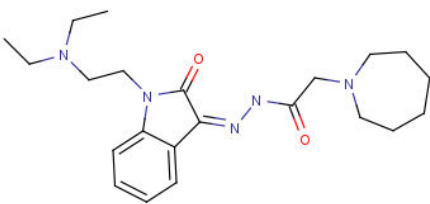
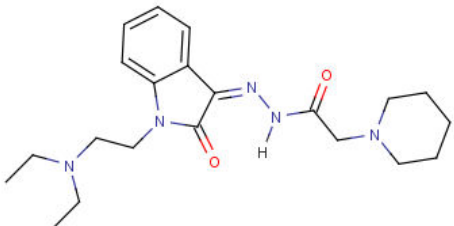
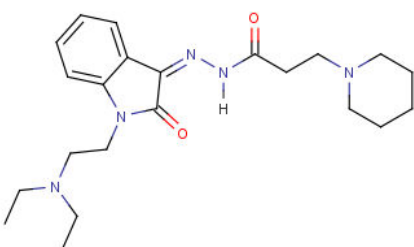
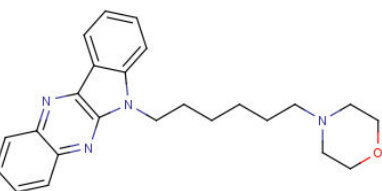
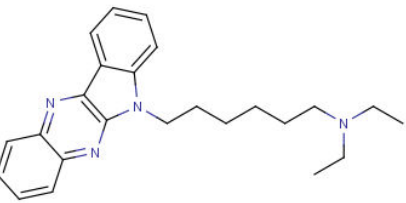
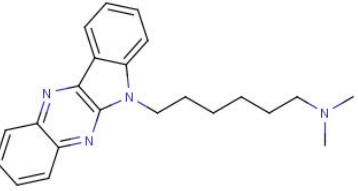
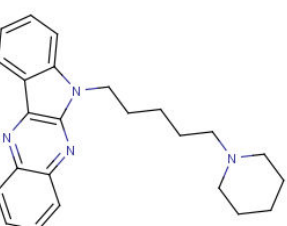
	65	5.07	30
	66	5.87	5
	67	5.2	50
	68	5.03	75
	69	4.93	85
	70	5.32	60
	71	5.92	

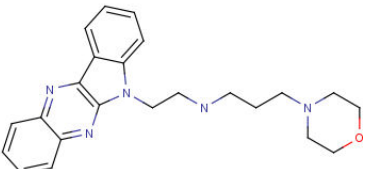
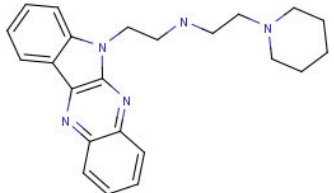
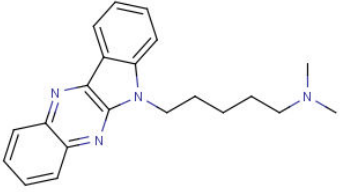

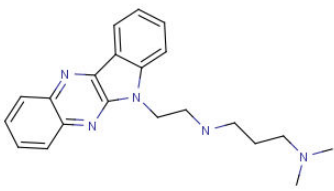
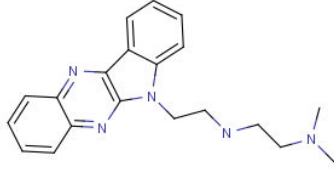
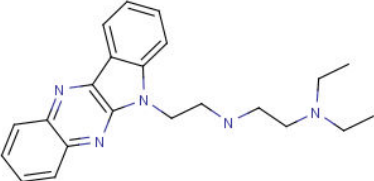
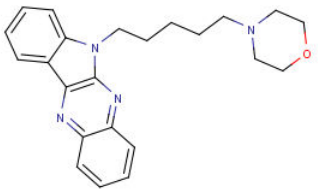
	72	5.79	10
	73	5.13	50
	74	5.15	80
	75	5.38	82
	76	5.72	
	77	5.37	70
	78	6.22	10

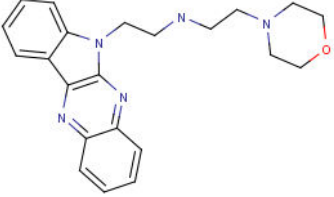
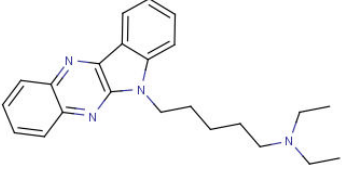
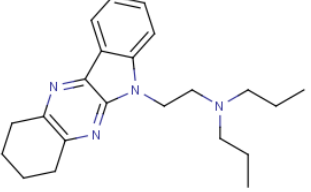
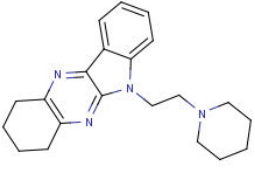
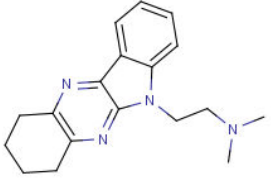
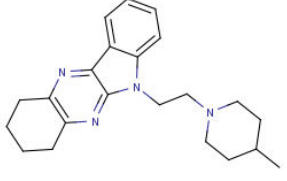
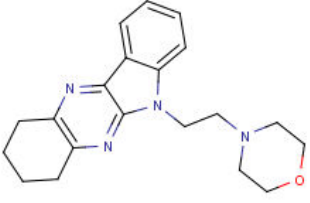
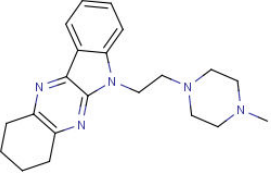
	79	5.49	13
	80	5.8	20
	81	5.94	35
	82	5.99	0
	83	6.03	
	84	5.92	0
	85	5.91	80
	86	6.15	100

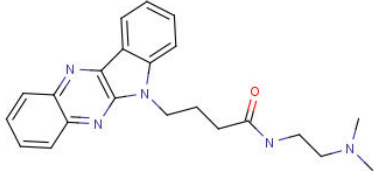
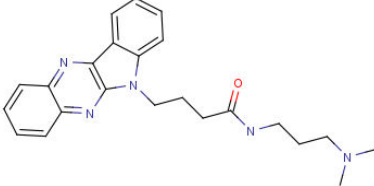
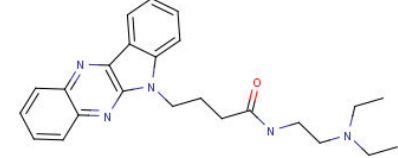
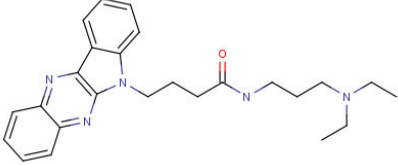
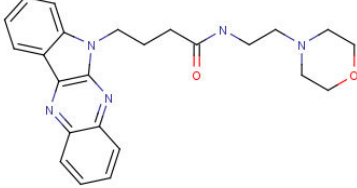
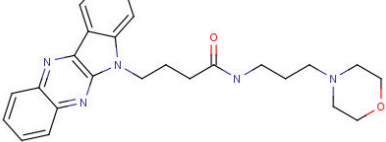
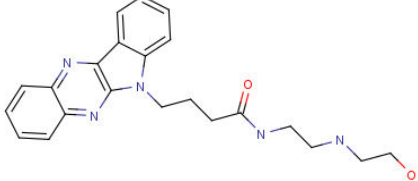
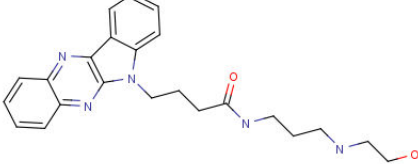
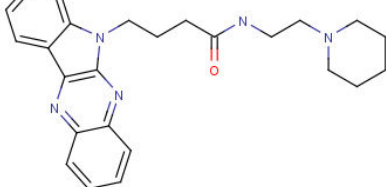
	87	6.47	100
	88	6.01	23
	89	6.11	30
	90	6.13	41
	91	6.04	28
	92	6.07	74
	93	6.28	56
	94	6.37	39

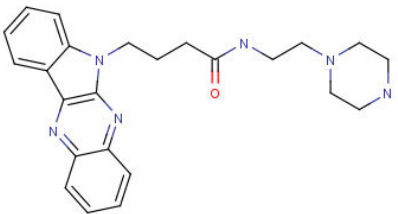
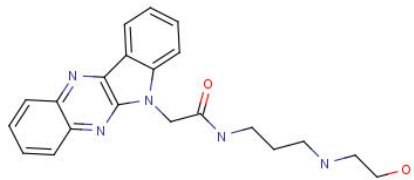
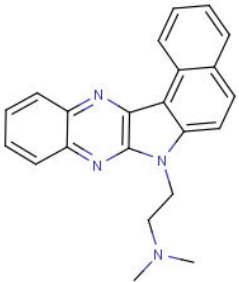
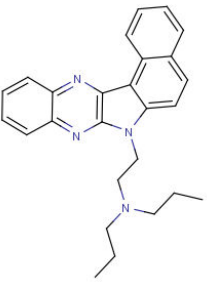
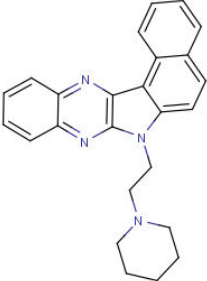
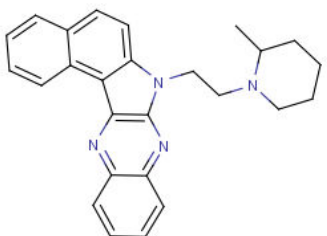
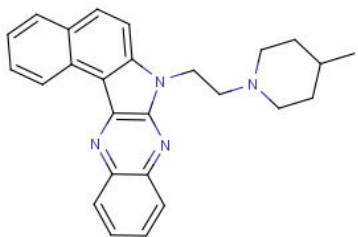
	95	6.09	25
	96	6.4	37
	97	6.08	
	98	6.31	80
	99	5.61	30
	100	5.93	11
	101	5.6	26
	102	6.04	10

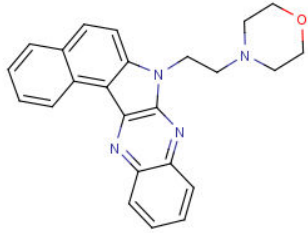
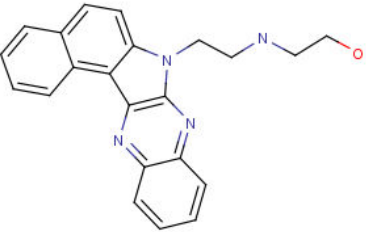
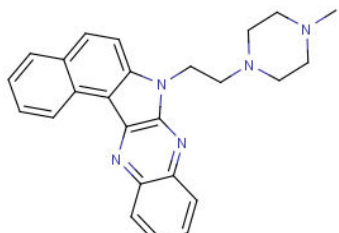
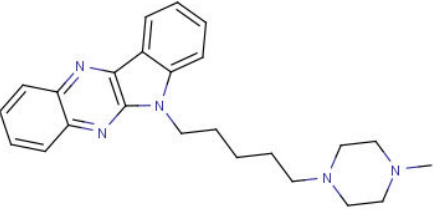
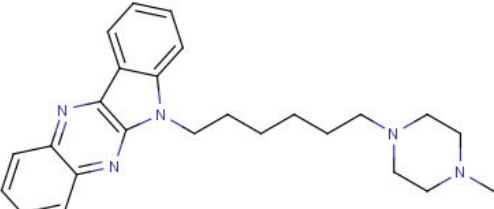
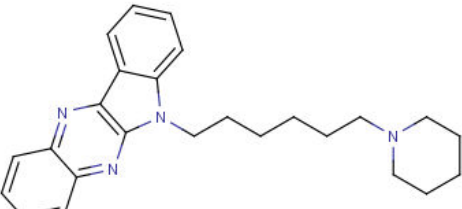

	103	5.96	32
	104	6.32	
	105	5.74	
	106	5.39	100
	107	5.73	6
	108	5.65	75
	109	5.81	82
	110	5.68	89

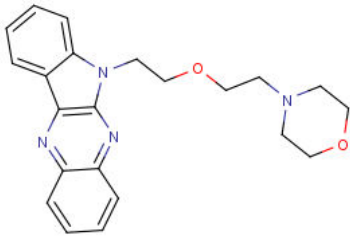
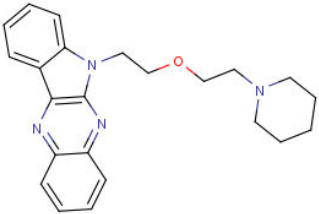
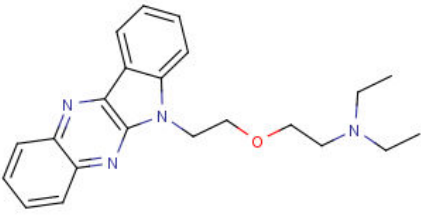
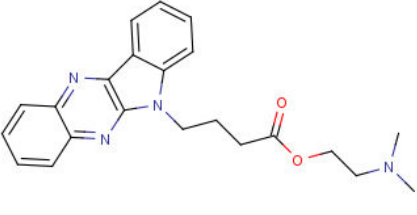
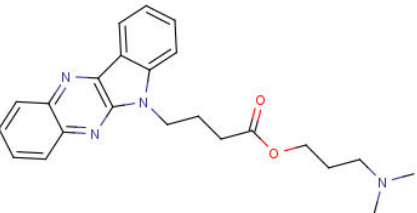
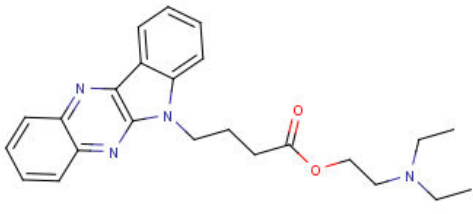
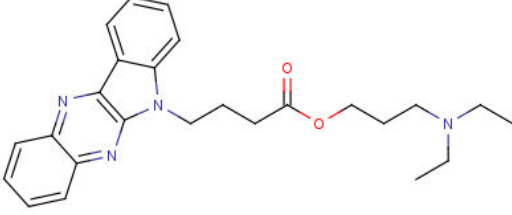
	111	6.72	52
	112	6.85	52
	113	5.66	68
	114	6.87	53
	115	6.76	69
	116	6.83	42
	117	6.82	73
	118	5.73	77

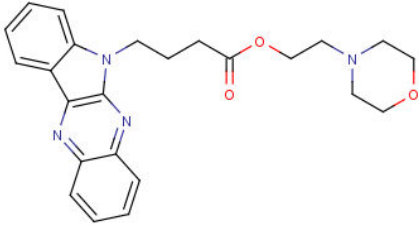
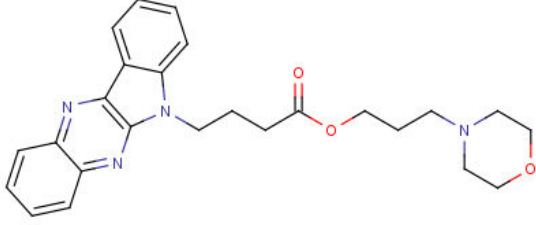
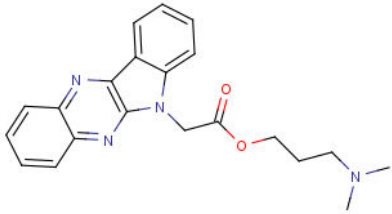
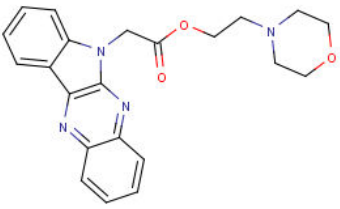
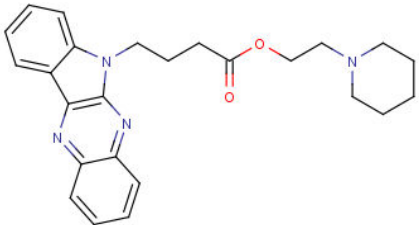

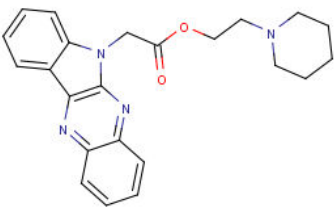
	119	6.17	40
	120	5.61	12
	121		90
	122		65
	123		88
	124		95
	125		66
	126		100

	127	5.01	
	128	5.09	
	129	5.25	
	130	4.77	
	131	4.91	
	132	5.11	
	133	5.05	
	134	5.09	
	135	4.98	

	136	4.97	
	137	5.06	72
	138	6.94	21
	139	6.91	16
	140	6.61	20
	141	6.83	20
	142	6.83	30

	143	6.94	10
	144	6.8	20
	145	6.58	17
	146	6.23	
	147	6.45	
	148	5.95	
	149	5.59	32

	150	5.32	50
	151	5.19	51
	152	5.26	56
	153	5.56	75
	154	5.32	100
	155	5.45	0
	156	5.29	0

	157	5.7	80
	158	5.35	100
	159	5.28	100
	160	5.18	45
	161	5.57	25
	162	5.46	0
	163	5.39	50

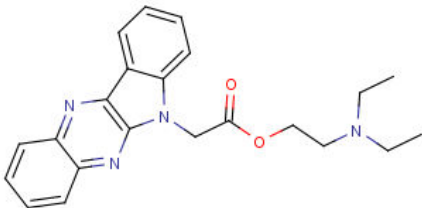
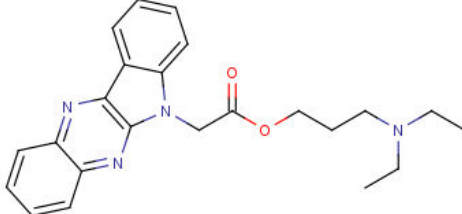
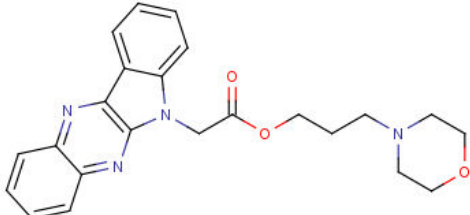
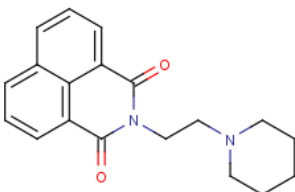
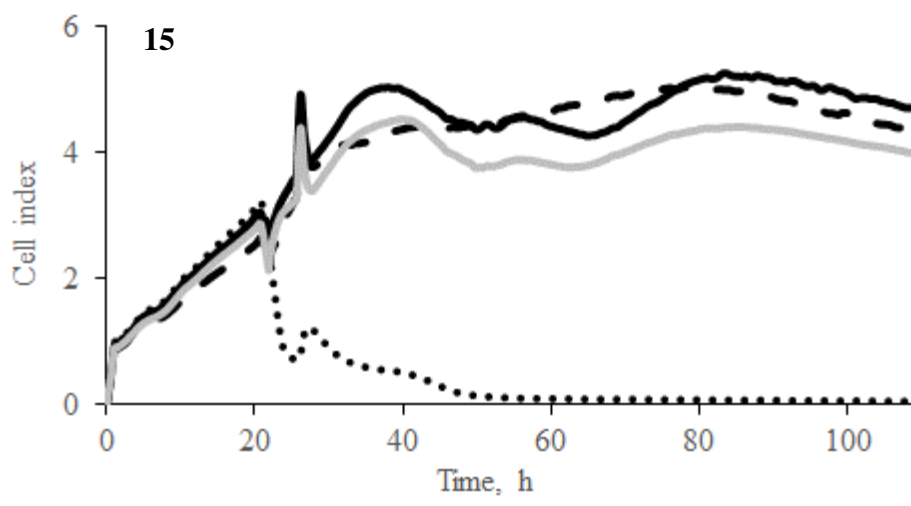
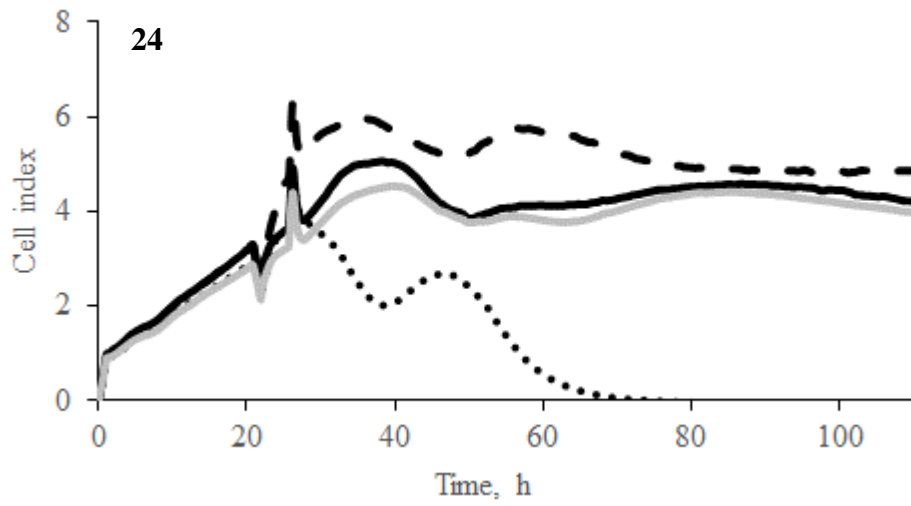
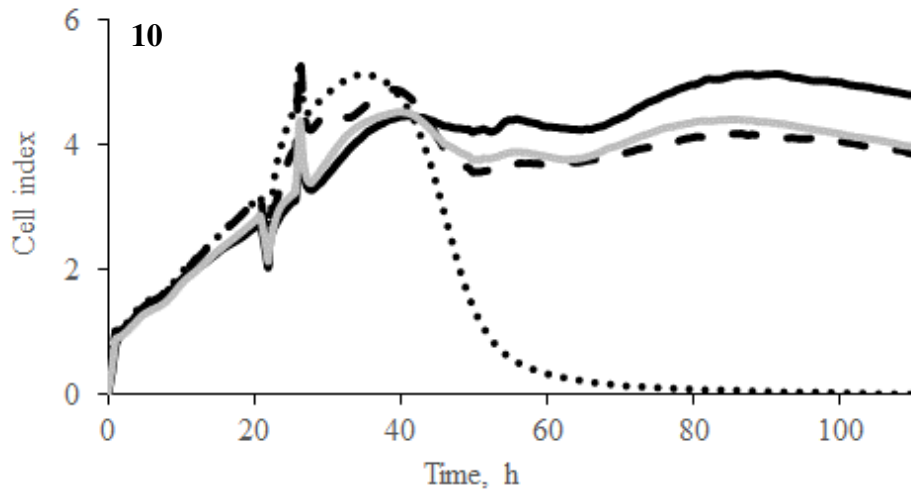
	164	5.42	45
	165	5.17	0
	166	5.14	15
	167	5.29	

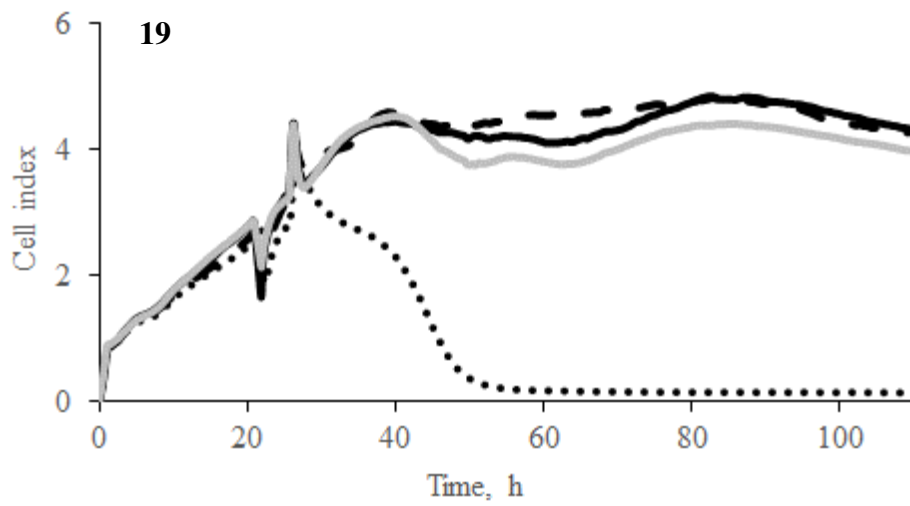
Table S5 All results of cytotoxicity and antiviral activity measured by Flow cytometry analysis

Compound ID	Cells viability, %	CC50, μ M	Relative expression of GFP, %
K+			100
K-			8
4	62 \pm 4	\approx 10	67
5	60 \pm 5	100	91
8	82 \pm 8	\approx 50	91
9	87 \pm 6	100	111
10	79 \pm 10	\approx 100	23
11	92 \pm 10	>100	77
12	78 \pm 10	\approx 100	91
14	99 \pm 7	>100	77
15	88 \pm 9	\approx 50	45
16	99 \pm 8	>100	91
17	74 \pm 8	\approx 50	28
19	79 \pm 6	\approx 50	48
20	113 \pm 14	>100	71
23	94 \pm 10	>100	59
24	84 \pm 16	100	42
25	100 \pm 12	>100	67
26	88 \pm 4	\approx 100	83
27	109 \pm 13	>100	125
28	102 \pm 17	>100	83
30	93 \pm 7	\approx 100	100
31	96 \pm 7	>100	111
32	88 \pm 8	100	91
34	95 \pm 8	>100	111
35	85 \pm 5	>100	111
36	93 \pm 8	\approx 100	100
38	87 \pm 19	\approx 100	83
43	102 \pm 11	>100	71
44	94 \pm 11	100	125

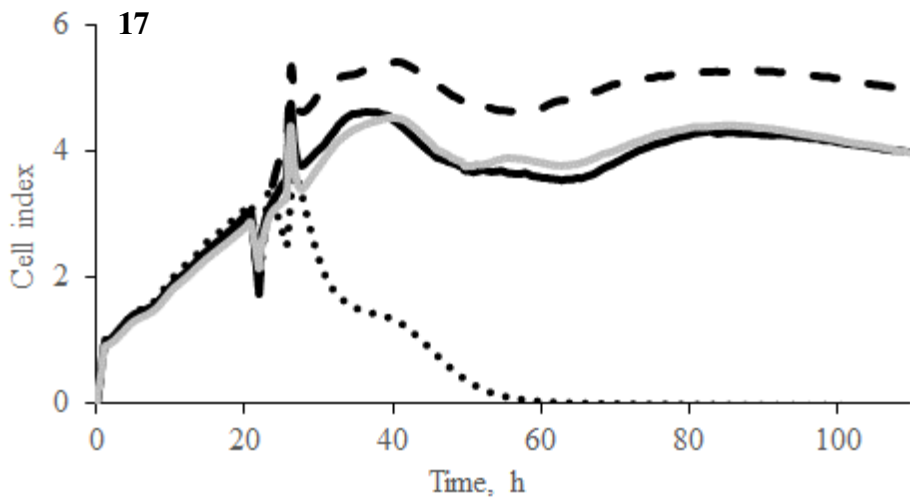
47	102±8	>100	111
50	65±12	100	77
51	78±13	>100	77
52	90±7	>100	83
53	88±7	>100	83
54	79±6	100	71
55	70±6	>100	83
58	86±16	≈50	111
60	84±6	≈100	48
61	99±17	>100	111
62	97±17	>100	111
63	104±20	>100	125

Figure S1 Cells viability within 110 h in the presence of 10, 24, 15, 19, 17 at different concentration.





— 1 μM — 10 μM
 100 μM — untreated cells



— 1 μM - - - 10 μM
 50 μM — untreated cells

APPENDIX C Published and submitted articles

Novel Enhanced Applications of QSPR Models: Temperature Dependence of Aqueous Solubility

Kyrylo Klimenko,^[a,b] Victor Kuz'min,^[a] Liudmila Ognichenko,^[a] Leonid Gorb,^[c] Manoj Shukla,^[d] Natalia Vinas,^[d] Edward Perkins,^[d] Pavel Polishchuk,^[e] Anatoly Artemenko,^[a] and Jerzy Leszczynski^{*,[f]}

A model developed to predict aqueous solubility at different temperatures has been proposed based on quantitative structure–property relationships (QSPR) methodology. The prediction consists of two steps. The first one predicts the value of k parameter in the linear equation $\lg S_w = kT + c$, where S_w is the value of solubility and T is the value of temperature. The second step uses Random Forest technique to create high-efficiency QSPR model. The performance of the model is

assessed using cross-validation and external test set prediction. Predictive capacity of developed model is compared with COSMO-RS approximation, which has quantum chemical and thermodynamic foundations. The comparison shows slightly better prediction ability for the QSPR model presented in this publication. © 2016 Wiley Periodicals, Inc.

DOI: 10.1002/jcc.24424

Introduction

The aqueous solubility (S_w) of organic compounds is a physico-chemical property, which is widely used in several scientific disciplines, such as chemistry, biology (including pharmaceutical area), and environmental science. Accurate experimental determination of aqueous solubility by shake-flask or turbidimetric methods is quite difficult, expensive, and time consuming, especially when solubility is low and compounds are potentially hazardous. A very good review on experimental approaches to determine aqueous solubility can be found in Ref. [1].

An attractive option is to predict aqueous solubility computationally. This could be done by using various approaches, based on different methodologies. For instance, quantitative structure–property relationships (QSPR) methodology can be used for regression and classification models development for aqueous solubility prediction based on structure-dependent variables,^[2] whereas other approaches use quantum mechanics or classical thermodynamics for solubility prediction.^[3,4] There are also some mixed models which use the information obtained at quantum-chemical level to build QSPR models (e.g., see Ref. [5]).

Recently, we have investigated aqueous solubility of military-relevant compounds using QSPR approach.^[6,7] In addition, QSPR analysis of aqueous solubility of more than 2500 organic compounds which belong to different classes and the influence of salinity on solubility was the subject of other publications.^[8,9] However, all models available in the literature for aqueous solubility prediction at QSPR level suffer from a serious limitation. They predict solubility in a quite narrow temperature range (typically 20–30°C),^[10–13] even though it is well known that solubility is a temperature-dependent property. For instance, solubility of nitroaromatic compounds which is of our scientific interest increases more than five times in the

range from 5 to 40°C.^[9] Such strong temperature dependence plays critical role in technological processes of industrial chemistry, drug design, or environmental sciences. Therefore, in the current work, we have broadened the principles and techniques formulated in Ref. [8] by adding one additional parameter—the temperature of dissolution inclusion into QSPR consideration.

There was an attempt to develop QSPR model to predict solubility for single class of substances (anthraquinones) at different temperatures. In this study,^[14] temperature was used as a molecular descriptor in Wavelength Neural Network model for prediction of a solubility of 25 anthraquinone dyes in supercritical carbon dioxide at 18–150°C. In our case, models are based on several hundred organic compounds of various classes, and to the best of our knowledge, this is the first time that temperature parameter has been directly incorporated in

[a] K. Klimenko, V. Kuz'min, L. Ognichenko, A. Artemenko
Department of Molecular Structure and Chemoinformatics, A.V. Bogatsky
Physical-Chemical Institute National Academy of Sciences of Ukraine,
Lustdorfskaya Doroga 86, Odessa 65080, Ukraine

[b] K. Klimenko
Laboratoire de Chemoinformatique, (UMR 7140 CNRS/UniStra) Université
de Strasbourg, 1, rue B. Pascal, Strasbourg 67000, France

[c] L. Gorb
HX5 LLC, Vicksburg, Mississippi 39180

[d] M. Shukla, N. Vinas, E. Perkins
US Army Engineer Research and Development Center, Vicksburg,
Mississippi 39180

[e] P. Polishchuk
Institute of Molecular and Translational Medicine, Palacky University
Olomouc, Hnevotínská 1333/5, Olomouc 779 00, Czech Republic

[f] J. Leszczynski
Interdisciplinary Center for Nanotoxicity, Department of Chemistry, Jackson
State University, Jackson, Mississippi 39217
E-mail: jerzy@icnanotox.org

© 2016 Wiley Periodicals, Inc.

the QSPR model that predicts aqueous solubility for compounds of different classes.

Materials and Methods

Recent reviews^[2,8] provide the state-of-art analysis of QSPR applicability in terms of aqueous solubility description and prediction. It was highlighted that current tendency in QSPR models development is to create models capable of describing and predicting solubility of large sets of structurally diverse compounds. As solubility is rather not an additive property, the choice of QSPR descriptors also has to be specific. Such specificity can be assured using Simplex Representation of Molecular Structure (SiRMS) approach.^[6,8] This method to generate QSAR descriptors has been described briefly here.^[15,16] At the 2D level, the connectivity of atoms in a fragment, atom type, and bond nature (single, double, triple, or aromatic) is taken into account. SiRMS approach accounts not only for the atom type but also for other atomic characteristics that may impact the physical and chemical properties of molecules, for example, partial charge, lipophilicity, refraction, and the ability of an atom to be a donor/acceptor in hydrogen-bond formation (H-bond). For atom characteristics with continuous values (i.e., charge, lipophilicity, and refraction), the subdivision of the entire value range into discrete groups was suggested. The values of these properties are calculated for every atom in the molecule following Jolly-Perry algorithm of electronegativity equalization^[17] for partial atom charges, XlogP scheme for lipophilicity,^[18] and the atomic refraction scheme suggested by Ioffe.^[19] Then, the atoms have been divided into four groups corresponding to their (i) partial charge ($A \leq -0.05 < B \leq 0 < C \leq 0.05 < D$), (ii) lipophilicity ($A \leq -0.5 < B \leq 0 < C \leq 0.5 < D$), (iii) refraction ($A \leq 1.5 < B \leq 3 < C \leq 8 < D$), (iv) van der Waals attraction ($50 < 100 < 250 < 400 < 650 < 2000$), and (v) van der Waals repulsion ($20,000 < 32,000 < 50,000 < 100,000$). For H-bond characteristics, the atoms have been divided into three groups: A (acceptor of hydrogen in H-bond), D (donor of hydrogen in H-bond), and I (indifferent atom). This algorithm is implemented in HiT QSAR software 6 which was used in this study. A more detailed information on description types is given in the Descriptor type section in Supporting Information.

QSPR model development was carried out using Random Forest (RF; Ref. [20]) statistical approach. RF method is based on decision tree algorithm, particularly growing decision trees in ensembles and then allowing them voting for the most popular class. Recent advances made it possible to use RF for accurate prediction of numerical data. This makes RF an effective nonparametric statistical technique for large database analysis. The main features of RF are listed below:

1. there is no need for descriptors preselection;
2. analysis of compounds with different mechanism of action within one dataset;
3. the method has its own out-of-bag procedure for the estimation of model quality and its internal predictive ability; and

4. models obtained are tolerant to "noise" in source experimental data.

Out-of-bag means when each new training set is drawn, with replacement, from the original training set. Then, a tree is grown on the new training set using random feature selection. Given a specific training set, form bootstrap training sets, construct classifiers $h[x, \text{Training}(\text{bootstrap})]$, and let these vote to form the bagged predictor. For each y, x in the training set, aggregate the votes only over those classifiers for which bootstrap training set does not contain y and x .^[20] This procedure guarantees that every compound from the training set will be in the internal test set at least once. Another internal prediction capacity assessment was carried out using n -fold cross-validation. In the n -fold cross-validation,^[21] sometimes called rotation estimation, dataset is randomly split into n mutually exclusive subsets of approximately equal size. Quantitative characteristics of the performance of models were assessed using determination coefficient (R^2) and root-mean-square error (RMSE).

The RF approach has not been widely used for QSPR studies yet^[22–25]; however, RF methodology proves to be very useful in our recent "structure–aqueous solubility" investigation.^[8] The CF^[26] software was used to perform model development.

Dataset

The data on aqueous solubility of a large set of compounds at different temperatures were taken from Ref. [27]. However, not all 4661 compounds from the handbook were used in model development. To perform the assessment of data accuracy, we applied the data evaluation system presented in Ref. [27]. It consists of five parameters (temperature, purity of solute, equilibrium time/agitation, analysis, and accuracy and/or precision) evaluated using three grades: 0, 1, and 2 (low, medium, and high). Only compounds which had temperature, purity of solute, and accuracy and/or precision criteria assessed at least as medium were chosen for further study. Second, some classes of organic compounds (namely, organic salts, polymeric compounds, and crystalline hydrates) were excluded from data set due to difficulties of their representation by molecular descriptors. In addition, it was crucial to remove mixtures, duplicates, and compounds with ambiguous CAS number.

As a result, 1484 aqueous solubility data points in the temperature range 4–97°C for 562 organic compounds have been used to form data set. It comprises compounds from various classes used in medicine (e.g., barbituric acid and benzodiazepine derivatives), agriculture (e.g., thiophosphate pesticides), and military (e.g., nitroaromatics). Solubility data were described as mole per liter with data points dispersed between -11.9 and $1.18 \lg(\text{mol/l})$ units. Among members of the data set, 141 compounds have solubility data for at least three temperatures, which allowed determining solubility–temperature curves needed for solubility temperature coefficient determination.

Compound CAS numbers and experimental and calculated $\lg S_w$ values at given temperature are given in Supporting Information Table S1.



Figure 1. The best fitting equations and maximum number of compounds for which temperature dependence of solubility is described appropriately according to F -test value. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

Workflow

First step of our study is to find a way to include the temperature of dissolution as the additional descriptor to expand existing descriptors defined in the section “Applied Methods and Models.” Similar to other processes, the process of dissolution follows the second principle of thermodynamics.^[28]

$$-RT \ln K = \Delta H - T\Delta S, \quad (1)$$

and Van't Hoff's equation:

$$\frac{d \ln K}{dT} = \frac{\Delta H}{RT^2}, \quad (2)$$

The application of Van't Hoff's equation (3) to the process of dissolution^[29] results in the following:

$$\ln x = \frac{\Delta H_{\text{fus}}}{R(T_m - T)} \quad (3)$$

where x is the mole fraction of solubility, T_m is the temperature of melting point, T is the temperature of dissolution, and ΔH_{fus} is the enthalpy of fusion.

As ΔH_{fus} for the most of crystalline compounds is positive, the solubility will be increased with the increasing of dissolution temperature. However, eq. (3) is not very popular in those applications that predict solubility, as in many cases, the values of melting point and enthalpy of fusion are not available. Therefore, semiempirical equations are suggested to predict solubility. Apelblat equation^[30–32] is one of the popular equations and is given by the following formula:

$$\ln x_{\text{eq}} = \frac{A}{T} + B + CT, \quad (4)$$

where x_{eq} is the mass ratio, T is the temperature of dissolution, and A , B , and C are regression coefficients.

The temperature of dissolution as a QSPR descriptor in this work was introduced as follows: First, the empirical equation was defined to fit the wide range of solubility data with acceptable accuracy. To achieve this goal, initially, we selected 18 compounds with solubility data points ranging from 4 to 11 from the dataset of 562 compounds. Thus, the selected range covers the data points ranging from very large aqueous solubility ($\log S = 0.9$) to practically insoluble substances ($\log S = -11.89$).

To define empirical equation which will be used further in our QSPR study, the discussed dataset has been treated by

TableCurve 2D software^[33] to perform regression analysis and to determine the empirical equation fitting the temperature dependence of solubility data in the best possible manner. The results of analysis are presented in Figure 1. In addition, Supporting Information Table S2 collects the F -test values.

There are two equations which have fitting temperature dependence of solubility data in the best way. They have the following forms:

$$\lg S_w = kT + c, \quad (5)$$

and

$$\lg S_w = \frac{kT}{\ln T} + c. \quad (6)$$

We would also like to mention that eq. (4) does not show high F -test value for any of the 18 compounds (see data shown in Supporting Information Table S2) and therefore will not be used for further QSPR study.

Linear equation (5) has the simplest form and it describes that usually temperature rise will lead to increase in solubility. Therefore, we decided to use this equation for QSPR analysis in our study.

However, the performed regression analysis also revealed that in case of eq. (5), the regression coefficient k is not a constant value (it varies from -6.30×10^{-4} to 3.35×10^{-4} among 18 compounds), and therefore, the value of temperature by itself as molecular descriptor is insufficient. In other words, we found that in eq. (5), suitable QSPR descriptor that describes temperature dependence of solubility is kT but not T by itself. However, most of the 562 compounds from the data set do not have enough data to derive that coefficient for every compound simply by solving the equation using the least squares method.

To overcome this issue, it was decided to use inductive transfer approach for descriptor calculations.^[34] In the framework of this approach, the individual models are not viewed as separate prediction tools but as nodes in the network of mutually dependent models built in parallel by means of multitask learning, or sequentially, using feature nets (FNs). FN uses extra tasks to build the model, prediction of which are further used as extra inputs for the main task.^[34] FN represents a kind of sequential inductive transfer: the models for the main task are built using the results of auxiliary tasks models. In our case, it means that a separate QSPR model for the prediction of regression coefficients k has to be developed. Afterward, molecular descriptors which comprise both the

Table 1. Statistical parameters for temperature term (k) QSPR model.

Fold	Tree count	Variable count	Training set				Test set		
			N	R^2	$R^2(\text{oob})$	RMSE	n	R^2	RMSE
1	150	50	113	0.97	0.75	0.026	28	0.81	0.065
2	150	70	113	0.98	0.76	0.065	28	0.61	0.087
3	150	50	113	0.97	0.77	0.026	28	0.85	0.055
4	150	50	113	0.97	0.74	0.027	28	0.83	0.055
5	250	70	113	0.97	0.73	0.027	28	0.81	0.064
Average				0.97	0.75	0.034		0.78	0.066

temperature and regression coefficient will be used for the development of QSPR model for solubility prediction.

To create a QSPR model that is able to predict k , the following approaches have been used:

1. The data set on the solubility of 18 compounds has been augmented by the data of additional 123 compounds (see Supporting Information Table S3) that also have three or more data points on temperature dependence of solubility.
2. The k values have been calculated manually for all 141 compounds of this data set using eq. (5).
3. QSPR model based on SiRMS descriptors, the values of kT , and solubility values has been created. Because of the fact that temperature impact on solubility is rather low in small ranges, the T values have been divided by 10 and the k value was transformed into cubic root of k for scaling purposes. The temperature was incorporated into model as a molecular descriptor $\sqrt[3]{k[(T - 20)/10]}$.

To assure that our approach is not overcomplicated, we built a model which includes temperature parameter $\sqrt[3]{k[(T - 20)/10]}$, as well as models which include T , $1/T$, T^2 , and $1/T^2$ and compare models' predictive capacities. Finally, lipophilicity molecular descriptor was included,^[35] and the predicted results were inserted into model using FN technique.

The obtained results are presented in Table 1 along with Supporting Information Table S3 which shows the values of estimated $\sqrt[3]{k}$ coefficients. The performance of the model was assessed using fivefold cross-validation, which means that ~20% of solubility values were selected for every fold's internal test set as well as in-built out-of-bag method for RF.^[20]

Considering the fact that temperature solubility coefficient is initially a calculated parameter, it would be hard to expect high RMSE value for training and test set as there is no data on experimental errors. However, further water solubility model development has shown that the quality of calculated temperature solubility coefficients is good enough to create a powerful tool for aqueous solubility prediction.

Results

We extended the above discussed QSPR approach to the rest of the 421 organic compounds to evaluate temperature dependence of aquatic solubility. The computed results collected in Table 2 show that such a model has high R^2 values for both training and test sets. Taking into account of the fact that considered compounds were structurally diverse and no standard procedure for experimental solubility determination was carried out, RMSE value can be regarded as quite acceptable. This value is comparable with the one obtained from experimental measurement of the solubility where it is equal to 0.24 log units.^[36]

Even though cross-validation is a powerful tool for evaluating the model's quality, it was decided to use an external test set for model validation as well. Therefore, five compounds which were not used in previous training and test sets with solubility at different temperatures obtained from different sources^[37-41] were selected for this purpose.

Table 2. Random Forest statistical results for temperature dependence of water solubility QSPR modeling.

Fold	Tree count	Variable count	Training set				Test set		
			N	R^2	$R^2(\text{oob})$	RMSE	n	R^2	RMSE
1	200	150	1187	0.99	0.96	0.22	297	0.97	0.38
2	200	150	1187	0.99	0.96	0.22	297	0.97	0.35
3	200	150	1187	0.99	0.96	0.21	297	0.97	0.4
4	200	150	1187	0.99	0.96	0.21	297	0.96	0.41
5	200	150	1187	0.99	0.96	0.21	297	0.81	0.34
Average				0.99	0.96	0.21		0.78	0.38

The results of our predictions along with experimental data are presented in Table 3. For these five compounds, the comparison between experimental data and QSPR predicted ones results in fairly acceptable RMSE value equal to 0.77. An RF model which only used temperature as a descriptor had slightly worse predictive capacity on COSMO test set with RMSE of 0.38.

The last point of our testing is the comparison of predictive ability of our model with the capability of another computational approach that is able to predict a temperature dependence of solubility. For this purpose, we have selected the results of our recent study^[9] where we predicted the temperature dependence of solubility for nitro-compounds within COSMO-RS^[42] approach. As six compounds investigated in Ref. [9] are already included in our model's training set, we have selected the rest of four compounds that have in total 18 solubility data points for comparison. The results of such comparison are presented in the Table 3. As the RF model created has slightly better accuracy in predicting the solubility of above-mentioned compounds when compared with COSMO-RS approach [see RMSE (comparison) data placed in Table 3], we expect that it is slightly more accurate. Moreover, solubility values for these compounds were calculated within several seconds, which is not possible using quantum chemical calculations. In addition, in contrast to COSMO-RS data, the developed QSPR model shows the pattern of solubility similar to the experimental data. To illustrate this, we present the patterns of the solubility at 30°C.

COSMO-RS: 610-39-9 ~ 606-20-2 > 602-01-7 > 99-35-4

QSPR Predictions: 99-35-4 > 602-01-7 ~ 610-39-9 ~ 606-20-2

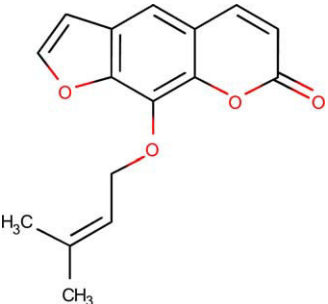
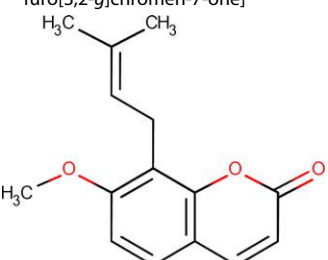
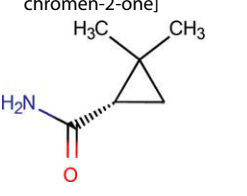
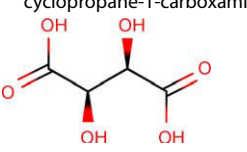
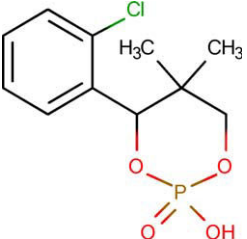
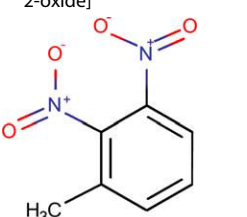
Experimental Data: 99-35-4 > 602-01-7 ~ 610-39-9 ~ 606-20-2

However, to be more conclusive in the comparison of the performance of RF and COSMO-RS methodologies, one needs to obtain similar temperature dependence values of water solubility for the same amount of compounds that has been considered for RF level. This is out of the scope of this particular work.


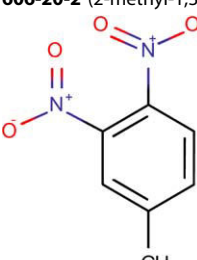

Conclusion

We have determined that the value of temperature of dissolution (T) by itself is not the best option of a descriptor that could be used in QSPR analysis to predict a temperature dependence of water solubility. Such a descriptor is the product between regression coefficient k of eq. (5) and the temperature of dissolution. Based on this analysis, we have developed

Table 3. External tests compounds prediction result. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

Structure, CAS number, and IUPAC name	T (°C)	Observation	Our model	COSMO
 <p>482-44-0 [9-(3-methylbut-2-enyloxy)furo[3,2-g]chromen-7-one]</p>	15	-4.39	-3.43	-
	20	-4.24	-3.41	-
	25	-4.23	-3.39	-
	30	-4.24	-3.38	-
	35	-4.12	-3.36	-
	40	-4.09	-3.36	-
	45	-4.08	-3.36	-
	50	-4.07	-3.35	-
	55	-4.00	-3.35	-
 <p>484-12-8 [7-methoxy-8-(3-methylbut-2-enyl)chromen-2-one]</p>	15	-4.61	-3.48	-
	20	-4.57	-3.47	-
	25	-4.52	-3.46	-
	30	-4.44	-3.44	-
	35	-4.40	-3.43	-
	40	-4.32	-3.43	-
	45	-4.28	-3.43	-
	50	-4.20	-3.43	-
	55	-4.13	-3.43	-
 <p>75885-58-4 [(S)-(+)-2,2-dimethylcyclopropane-1-carboxamide]</p>	22	-0.68	-0.36	-
	35	-0.68	-0.34	-
	42	-0.49	-0.33	-
	54	-0.40	-0.32	-
	59	-0.29	-0.31	-
	63	-0.19	-0.31	-
	68	-0.08	-0.30	-
	73	0.00	-0.30	-
	81	0.17	-0.29	-
 <p>87-69-4 [(2R,3R)-2,3-dihydroxybutanedioic acid]</p>	15	-0.60	0.31	-
	25	-0.54	0.42	-
	35	-0.44	0.47	-
	45	-0.38	0.47	-
	55	-0.33	0.47	-
	55	-0.33	0.47	-
	65	-0.29	0.47	-
 <p>98634-28-7 [4-(2-chlorophenyl)-2-hydroxy-5,5-dimethyl-1,3,2-dioxaphosphorinane 2-oxide]</p>	5	-2.16	-2.90	-
	10	-2.17	-2.90	-
	15	-2.12	-2.90	-
	20	-2.06	-2.89	-
	25	-2.05	-2.87	-
	30	-2.04	-2.87	-
	35	-2.04	-2.86	-
	40	-2.03	-2.85	-
	45	-2.02	-2.84	-
 <p>98634-28-7 [4-(2-chlorophenyl)-2-hydroxy-5,5-dimethyl-1,3,2-dioxaphosphorinane 2-oxide]</p>	5	-3.43	-2.9	-3.63
	7	-3.36	-2.9	-3.57
	19	-3.17	-2.87	-3.34
	30	-2.96	-2.82	-3.2
	41	-2.76	-2.79	-2.89

(Continued)

Table 3. (Continued)				
Structure, CAS number, and IUPAC name	T (°C)	Observation	Our model	COSMO
602-01-7 (1-methyl-2,3-dinitrobenzene) 	5	-3.39	-2.74	-3.58
	7	-3.31	-2.74	-3.54
	19	-3.09	-2.72	-3.25
	30	-2.84	-2.67	-2.99
	41	-2.63	-2.64	-2.74
606-20-2 (2-methyl-1,3-dinitrobenzene) 	5	-3.42	-2.9	-3.54
	20	-3.13	-2.87	-3.24
	31	-2.95	-2.81	-3.01
	40	-2.75	-2.79	-2.83
610-39-9 (4-methyl-1,2-dinitrobenzene) 	5	-2.9	-2.7	-4.21
	19	-2.74	-2.68	-3.9
	30	-2.58	-2.59	-3.66
	41	-2.45	-2.46	-3.41
99-35-4 (1,3,5-trinitrobenzene) RMSE (total) RMSE (comparison)			0.67	
			0.34	0.57

a twofold QSPR procedure to predict temperature dependence of water solubility of organic compounds. The first step uses SiRMS-generated descriptors to predict the value $\sqrt[3]{k}$. The second step applies both SiRMS-generated descriptors and a value of $\sqrt[3]{k[(T-20)/10]}$ to generate effective models that are able to accurately predict the temperature dependence of solubility. The successful predictive ability of these models has been illustrated by the application of independent external test set and the comparison with limited amount of the temperature-dependent water solubility values for the compounds obtained at COSMO-RS level.

Acknowledgments


The results in this study were funded and obtained from research conducted under the Environmental Quality Technology Program of the United States Army Corps of Engineers by the USAERDC. The findings of this report are not to be construed as an official Department of the Army position unless so designated by other authorized documents. Jerzy Leszczynski thanks the National Science Foundation for support from the NSF CREST Interdisciplinary Nanotoxicity Center (grant no. HRD-0833178; NSF-EPSCoR grant: 362492-190200-01/NSFEPS-0903787). The computation time was provided by the Extreme Science and Engineering Discovery Environment (XSEDE) by

the National Science Foundation (grant no.: OCI-1053575) and XSEDE award allocation number DMR110088 and by the Mississippi Center for Supercomputer Research. Authors are thankful to Natalia Sizochenko for her contribution in model development.

The use of trade, product, or firm names in this report is for descriptive purposes only and does not imply endorsement by the U.S. Government.

Keywords: QSPR · feature net · temperature-dependent · aqueous solubility

How to cite this article: K. Klimenko, V. Kuz'min, L. Ognichenko, L. Gorb, M. Shukla, N. Vinas, E. Perkins, P. Polishchuk, A. Artemenko, J. Leszczynski. *J. Comput. Chem.* **2016**, DOI: 10.1002/jcc.24424

 Additional Supporting Information may be found in the online version of this article.

- [1] A. Avdeef, In Absorption and Drug Development: Solubility, Permeability and Charge State; Wiley-Interscience, Hoboken, NJ, **2003**; Chapter 6, pp. 91–116.
- [2] I. V Tetko, In Computational Toxicology: Risk Assessment for Pharmaceutical and Environmental Chemicals; Wiley, Hoboken, NJ, **2007**; Part 3: The Prediction of Physicochemical Properties, pp. 241–277.

- [3] A. Klamt, F. Eckert, *Fluid Phase Equilib.* **2000**, *172*, 43.
- [4] A. Klamt, G. J. Schüürmann, *Chem. Soc. Perkin. Trans.* **1993**, *2*, 799.
- [5] O. Isayev, B. Rasulev, L. Gorb, J. Leszczynski, *Mol. Diversity* **2006**, *10*, 233.
- [6] E. N. Muratov, V. E. Kuz'min, A. G. Artemenko, N. A. Kovdienko, L. Gorb, F. Hill, J. Leszczynski, *Chemosphere* **2010**, *79*, 887.
- [7] Y. A. Kholod, E. N. Muratov, L. G. Gorb, F. C. Hill, A. G. Artemenko, V. E. Kuz'min, M. Qasim, J. Leszczynski, *Environ. Sci. Technol.* **2009**, *43*, 9208.
- [8] N. A. Kovdienko, P. G. Polishchuk, E. N. Muratov, A. G. Artemenko, V. E. Kuz'min, L. Gorb, F. Hill, J. Leszczynski, *Mol. Informatics* **2010**, *29*, 394.
- [9] Y. A. Kholod, G. Gryn'ova, L. Gorb, F. C. Hill, J. Leszczynski, *Chemosphere* **2011**, *83*, 287.
- [10] I. V. Tetko, V. Y. Tanchuk, T. N. Kasheva, A. E. P. Villa, *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1488.
- [11] D. Butina, J. M. R. Gola, *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 837.
- [12] A. Yan, J. Gasteiger, *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 429.
- [13] M. Salahinejad, T. C. Le, D. A. Winkler, *Mol. Pharm.* **2013**, *10*, 2757.
- [14] R. Tabaraki, T. Khayamian, A. A. Ensafi, *J. Mol. Graph. Model.* **2006**, *25*, 46.
- [15] V. Kuz'min, A. Artemenko, P. Polishchuk, E. Muratov, A. Hromov, A. Liahovskiy, S. Andronati, S. Makan, *J. Mol. Model* **2005**, *11*, 457.
- [16] V. E. Kuz'min, A. G. Artemenko, E. N. Muratov, *J. Comput.-Aided. Mol. Des.* **2008**, *22*, 403.
- [17] W. L. Jolly, W. B. Perry, *J. Am. Chem. Soc.* **1973**, *95*, 5442.
- [18] R. Wang, Y. Fu, L. Lai, *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 615.
- [19] B. V. Ioffe, *Chemistry of Refractometric Methods (in Russian)*; Khimia: Leningrad, **1983**; p. 350.
- [20] L. Breiman, *Mach. Learn.* **2001**, *45*, 5.
- [21] R. Kohavi, In Proceedings of the 14th International Joint Conference on Artificial Intelligence, Montreal, Quebec, Canada, August 20–25, 1995; Morgan Kaufmann Publishers: San Francisco, CA, USA, **1995**. Volume 2, p. 1137–1143.
- [22] V. Svetnik, A. Liaw, C. Tong, J. C. Culberson, R. P. Sheridan, B. P. Feuston, *J. Chem. Inf. Model* **2003**, *43*, 1947.
- [23] Z. Debeljak, A. Skrbo, et al. *J. Chem. Inf. Model.* **2007**, *47*, 918.
- [24] F. Lombardo, R. S. Obach, et al. *J. Med. Chem.* **2006**, *49*, 2262.
- [25] R. P. Sheridan, K. R. Korzekwa, R. A. Torres, M. J. Walker, *J. Med. Chem.* **2007**, *50*, 3173.
- [26] P. G. Polishchuk, E. N. Muratov, A. G. Artemenko, O. G. Kolumbin, N. N. Muratov, V. E. Kuz'min, *J. Chem. Inf. Model.* **2009**, *49*, 2481.
- [27] S. H. Yalkowsky, Y. He, In *Handbook of Aqueous Solubility Data*; CRC Press: Boca Raton, FL, **2003**; pp. 1–1366.
- [28] S. Walas, *Phase Equilibria in Chemical Engineering*; Butterworth-Heinemann, London, **1985**; Vol. 2, pp. 406–436.
- [29] F. L. Nordström, Å. C. Rasmuson, *Eur. J. Pharm. Sci.* **2009**, *36*, 330.
- [30] A. Apelblat, E. Manzurola, *J. Chem. Thermodyn.* **1997**, *29*, 1527.
- [31] A. Apelblat, E. Manzurola, *J. Chem. Thermodyn.* **1999**, *31*, 85.
- [32] E. Manzurola, A. Apelblat, *J. Chem. Thermodyn.* **2002**, *34*, 127.
- [33] TableCurve 2D version 5.01, trial version available from: <https://systatsoftware.com/products/tablecurve-2d/>.
- [34] A. Varnek, C. Gaudin, G. Marcou, I. Baskin, A. K. Pandey, I. V. Tetko, *J. Chem. Inf. Model.* **2009**, *49*, 133.
- [35] L. N. Ognichenko, V. E. Kuz'min, L. Gorb, F. C. Hill, A. G. Artemenko, P. G. Polishchuk, J. Leszczynski, *Mol. Inf.* **2012**, *31*, 273.
- [36] A. R. Katritzky, U. Maran, V. S. Lobanov, M. Karelson, *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1.
- [37] F. Sun, H. Kang, K. B. Liu, B. Zhang, *Fluid Phase Equilib.* **2012**, *330*, 12.
- [38] X. Shi, C. Zhou, Y. Gao, X. Chen, *Chin. J. Chem. Eng.* **2006**, *14*, 547.
- [39] W. Yang, K. Wang, Y. Hu, F. Shen, J. Feng, *J. Solution. Chem.* **2013**, *42*, 485.
- [40] G. D. Yang, C. Li, A. G. Zeng, Q. H. Qu, X. Yang, X. L. Bian, *Fluid Phase Equilib.* **2012**, *325*, 41.
- [41] G. D. Yang, C. Li, A. G. Zeng, Y. L. Guo, X. Yang, J. F. Xing, *J. Mol. Liq.* **2012**, *167*, 86.
- [42] A. Klamt, *J. Phys. Chem.* **1995**, *99*, 2224. DOI:10.1021/j100007a062

Received: 9 January 2016
Revised: 22 April 2016
Accepted: 17 May 2016
Published online on 00 Month 2016

Submitted to “Antiviral Research”

Computer-aided design, synthesis and biological evaluation of DNA intercalating antiviral agents

Kyrylo Klimenko,^[a,b] Sergei Lyakhov,^[b] Marina Shibinskaya,^[b] Alexander Karpenko,^[b] Gilles Marcou,^[a] Dragos Horvath,^[a] Marina Zenkova,^[c] Elena Goncharova,^[c] Rinat Amirkhanov,^[c] Andrei Krysko^[b], Sergei Andronati,^[b] Igor Levandovskiy^[f], Pavel Polishchuk,^[b,d] Victor Kuz'min,^[b] and Alexandre Varnek* ^[a,e]

^[a] K. Klimenko, Dr. G. Marcou, Dr. D. Horvath, Dr. A. Varnek
Laboratoire de Chimoinformatique, (UMR 7140 CNRS/UniStra)
Université de Strasbourg,
1, rue B. Pascal, Strasbourg 67000, France,
e-mail : varnek@unistra.fr

^[b] K. Klimenko, Dr. S. Lyakhov, Dr. M. Shibinskaya, Dr. A. Karpenko, Dr. A. Krysko, Dr. S. Andronati,
Dr. V. Kuz'min,
A.V. Bogatsky Physico-Chemical Institute of NAS of Ukraine,
Lyustdorfskaya doroga, 86, Odessa 65080, Ukraine

^[c] Dr. M. Zenkova, Dr. E. Goncharova, Dr. E. Amirkhanov,
Institute of Chemical Biology and Fundamental Medicine,
Siberian Branch of Russian Academy of Sciences,
8 Lavrentiev Avenue, Novosibirsk 630090, Russia

^[d] Dr. P. Polishchuk,
Institute of Molecular and Translational Medicine,
Palacky University Olomouc, Hněvotínská 1333/5,
Olomouc 779 00, Czech Republic

^[e] Dr. A. Varnek
Federal University of Kazan, Kremlevskaya str., 18, Kazan, Russia

^[f] Dr. I. Levandovskiy
Department of Organic Chemistry, Kiev Polytechnic Institute, Pr. Pobedy 37, 03056 Kiev, Ukraine.

Abstract: This paper describes computer-aided design of new anti-viral agents acting as DNA intercalators. Earlier obtained experimental data have been used to establish simple rules (structural filters), as well as, to build pharmacophore and QSAR models for selection of the most promising compounds. Virtual screening of databases containing more than 3M molecules resulted in 55 hits which were synthesized and tested for antiviral activity. Two compounds displaying high antiviral activity against *Vaccinia virus* and low cytotoxicity were recommended for further antiviral activity investigations.

Keywords: antiviral activity, vaccinia virus, structure-activity modelling, virtual screening, DNA affinity

Highlights:

- multi-stage virtual screening resulted in 55 new potential DNA/RNA intercalators
- two hits active against *Vaccinia virus* were found among 55 synthesized compounds
- both compounds were not cytotoxic, did not induce interferon levels and bound to DNA that supports the hypothesis about their intercalating mechanism

1. Introduction

Viral diseases have a severe negative impact on human life worldwide^[1,2] which motivates researchers to develop new antiviral drugs. Most of known target-specific antiviral compounds inhibit certain viral proteins, e.g. protease or polymerase^[3]. Such compounds are rather selective, have low toxicity and the reduced risk of adverse effects. Corresponding drug discovery projects are frequently supported by different cheminformatics tools. Thus, a combination of QSAR and docking methods were used to identify a novel influenza virus neuraminidase inhibitor which is more potent than commercialized drug Oseltamivir^[4]. Virtual screening workflow included similarity search, shape-based and pharmacophore models was used to discover HIV-1 reverse transcriptase dual inhibitors^[5]. Comprehensive virtual screening and multi-objective optimization strategy allowed to identify novel HIV-1 inhibitors with favourable pharmacokinetics profiles^[7].

Broad spectrum antiviral agents may, however, be more advantageous than target-specific compounds which may be effective to control emerging pathogens^[6]. There exist several major groups of broad-spectrum antivirals. One of them includes interferon and interferon inducers. Interferon is a protein produced as an immune response, inducing synthesis of protein kinase which phosphorylates initiation factor of translation and, therefore, prevents synthesis of viral proteins. The second group includes nucleotide analogs, i.e., substances which resemble DNA or RNA nucleotide but have an inappropriate nitrogenous base. Being captured by proteins or tRNA involved in the virus reproduction processes; they may lead to the synthesis of a non-coding sequences in viral nucleic acids.^[7] The third group includes nucleic acid intercalators which may entry between the parallel pairs of bases in double helix of DNA or RNA.^[8] To our knowledge, *in silico* approaches are rarely used in the design of broad spectrum antivirals and no computer-aided design of intercalators was reported so far.

This study is devoted to the computer-aided design of new nucleic acid intercalators displaying antiviral activity. Modern cheminformatics tools – Quantitative Structure-Activity Relationships (QSAR) and pharmacophore models – were used in virtual screening procedure in order to discover chemical structures of potential antivirals. Computationally selected compounds have been synthesized and tested for antiviral activity, which lead to discovery of two compounds highly active against *Vaccinia virus*. Computations, synthesis and biological tests are described in dedicated sections of the article, each regrouping both methodology description and main results. Some technical details are provided in the Supplementary Material.

2. Structure-activity modeling and virtual screening

The virtual screening workflow included three types of predictive models: structural filters, QSAR and pharmacophore models. All these models were built on a dataset of 167 compounds synthesized and tested at A.V. Bogatsky Physico-Chemical Institute^[14] (PCI dataset, see Table A1 in Supplementary Material). Each molecule contains a polycyclic planar fragment linked to basic amino group (Figure 1). According to type of polycyclic fragments, the dataset could be divided on seven classes of compounds, as it is shown in Figure 1. In this dataset, each compound was annotated by binding constants (K_i) measured by substitution of ethidium bromide in DNA^[15]. Another kind of biological activity - maximum antiviral effect E_{max} (%) within 0.2 - 620 μ M concentration range was measured for 117 compounds as described in^[14]

Structural filters represent simple rules aiming to select the compounds similar to those in the BCI dataset. Since intercalators must have a planar fragment to entry between two parallel nucleic base pairs and a side chain able to interact with phosphate residues via hydrogen bonding, the number of fused rings (FR), H-bond donors (HD), H-bond acceptors (HA) were used as filters parameters. The number of rotatable bonds (RB) reflecting molecular flexibility and molecular weight (MW) as a characteristics of molecular size were also considered. Analysis of structures in the BCI dataset performed with the ChemAxon IJC tool^[18] revealed the following parameters ranges: FR = 2 - 5, HD = 0 - 3, HA = 2-6, RB = 3-12 and MW = 268-443. Upon virtual screening procedure, the molecules having at least one of these parameters outside the specified ranges were filter out.

The pharmacophore model has been built with the LigandScout^[22] software on a subset of 161 compounds with $\lg(K_i) \geq 4$ recognized as reasonable DNA intercalators. ZINC database ^[23] was used as a source of decoys for validation of pharmacophore models. Three best pharmacophore models (Figure 3) having the highest recall and precision values (0.66-1.00) were selected for virtual screening workflow.

Classification model able to distinguish compounds with higher ($E_{\max} \geq 50\%$) and lower ($E_{\max} < 50\%$) maximum antiviral effect has been built using the Random Forest method^[26] and simplex descriptors^[24] Earlier, similar technique was successfully used in QSAR modeling of various antiviral activities.^{[25][20][21]} The model's applicability domain was assessed with Euclidean distance-based method.^[27] The model was trained on the training set containing 133 compounds randomly selected from the PCI dataset It well performs on the test set containing remaining 34 compounds (balanced accuracy BA = 0.78). More details about model development are given in the Section 2 of the Supplementary Materials.

Developed models have been applied to screen a dataset of 3 207 605 compounds composed from 3 207 317 compounds from the BioinfoDB^[17] database and 288 virtual compounds generated as a combination of scaffolds and some typical fragments from training set compounds (Figure 1). At the first step, the structural filters discarded the major part of compounds (Figure 2). Remaining 1 022 465 compounds were screened with the pharmacophore models retaining 884 compounds, see examples in Figure 3. At the next step, remaining compounds were screened first with the classification structure-activity model developed in this work, then with previously developed QSPR model ^[28] assessing aqueous solubility and with the PASS software ^[29] assessing affinities to the wide spectra of biological targets. Finally, 55 compounds displaying any side effects, toxicity and mutagenicity and for which predicted solubility in water was larger than 10^{-5} mol/l have been selected. The search of the PubChem ^[30] database revealed that no compounds among selected hits were previously used in antiviral bioassays. Detailed description of synthesized compounds is given in Section 4 of Supplementary Material.

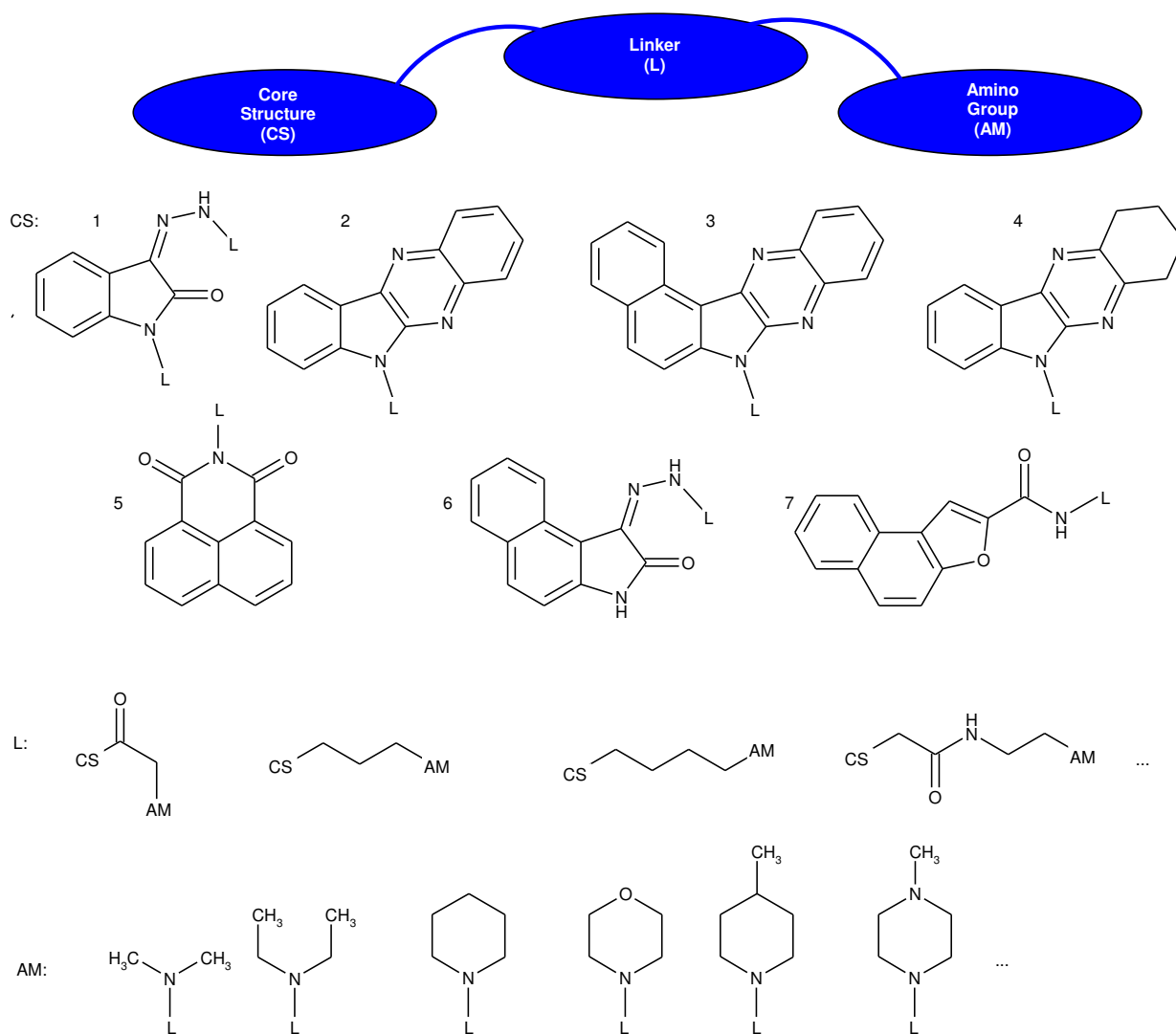


Figure 1 Seven classes of polycyclic molecules present in the modeling dataset

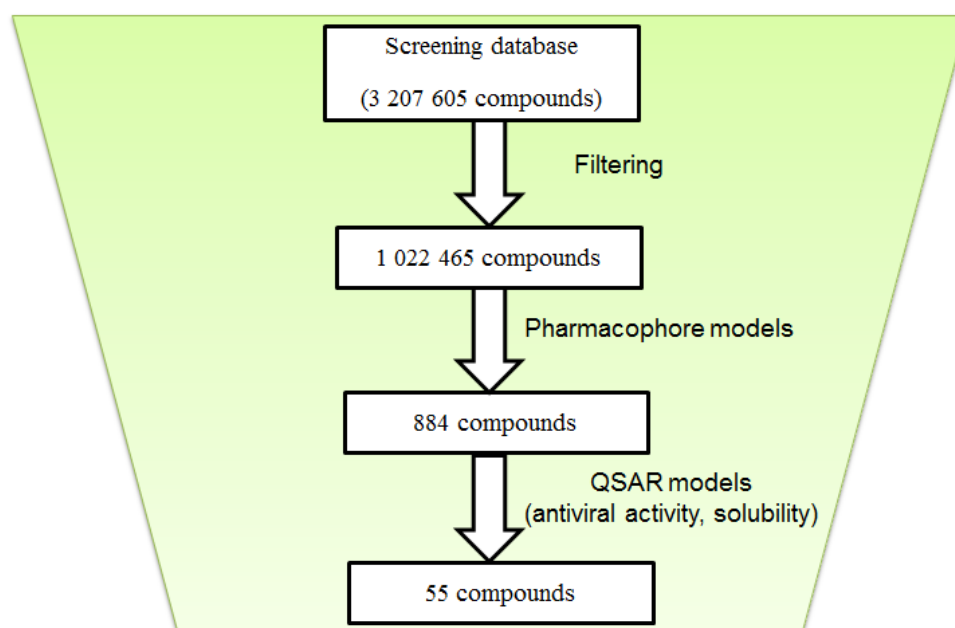


Figure 2. Virtual screening workflow.

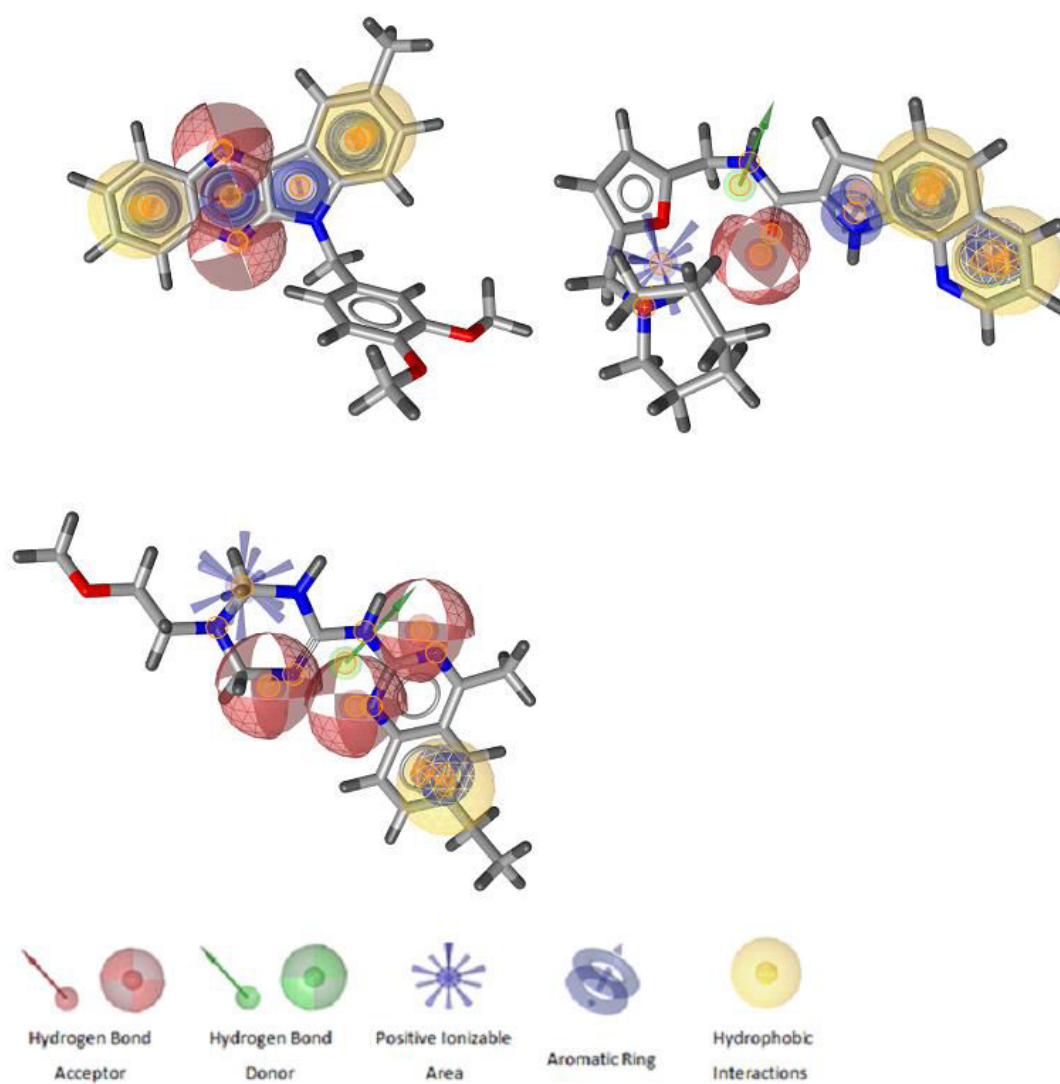
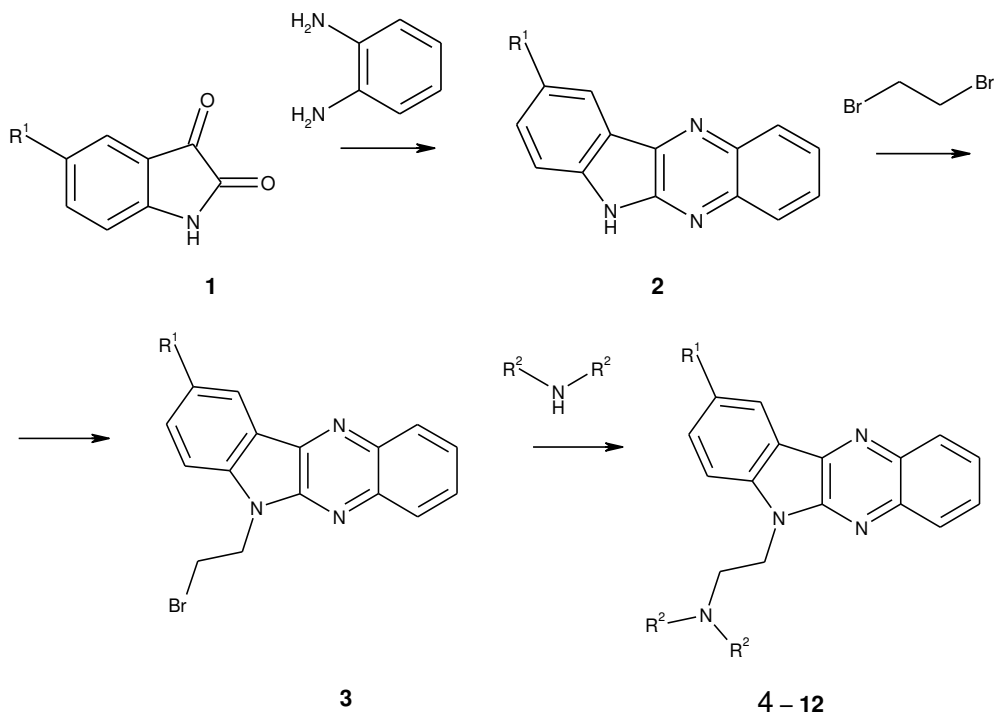


Figure 3. Some hits selected in virtual screening aligned with pharmacophore models. The description of pharmacophore features is given below.

3. Synthesis

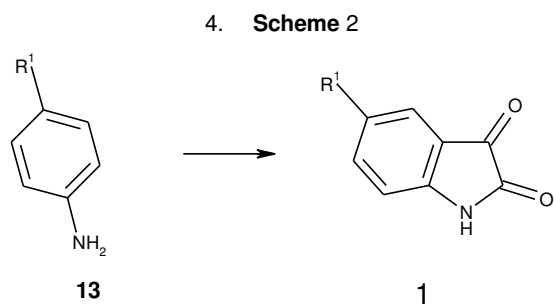
Compounds 4, 6, 11, 8, 12 were synthesized upon condensation of isatin (1, $R^1 = H$) with 1,2-diaminobenzene and consecutive alkylation (Scheme 1) as described in ^[14]

Scheme 1

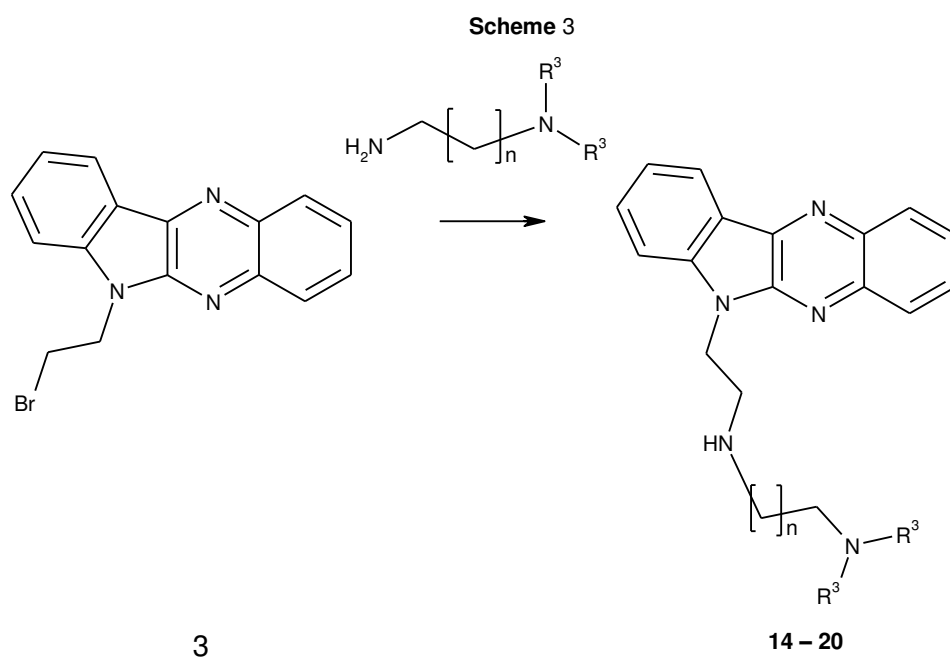


R^2-N-R^2	R^1		
	H	CH ₃	OCH ₃
	4	-	-
	-	-	5
	6	7	-
	8	9	-
	-	-	10
	11	-	-
	12	-	-

Notice that 5-substituted isatins **1** ($R^1 = \text{CH}_3, \text{OCH}_3$), precursors of compounds **5, 7, 9, 10** and were prepared from corresponding anilines **13** ($R^1 = \text{CH}_3, \text{OCH}_3$) by Sandmeyer's method (Scheme 2)

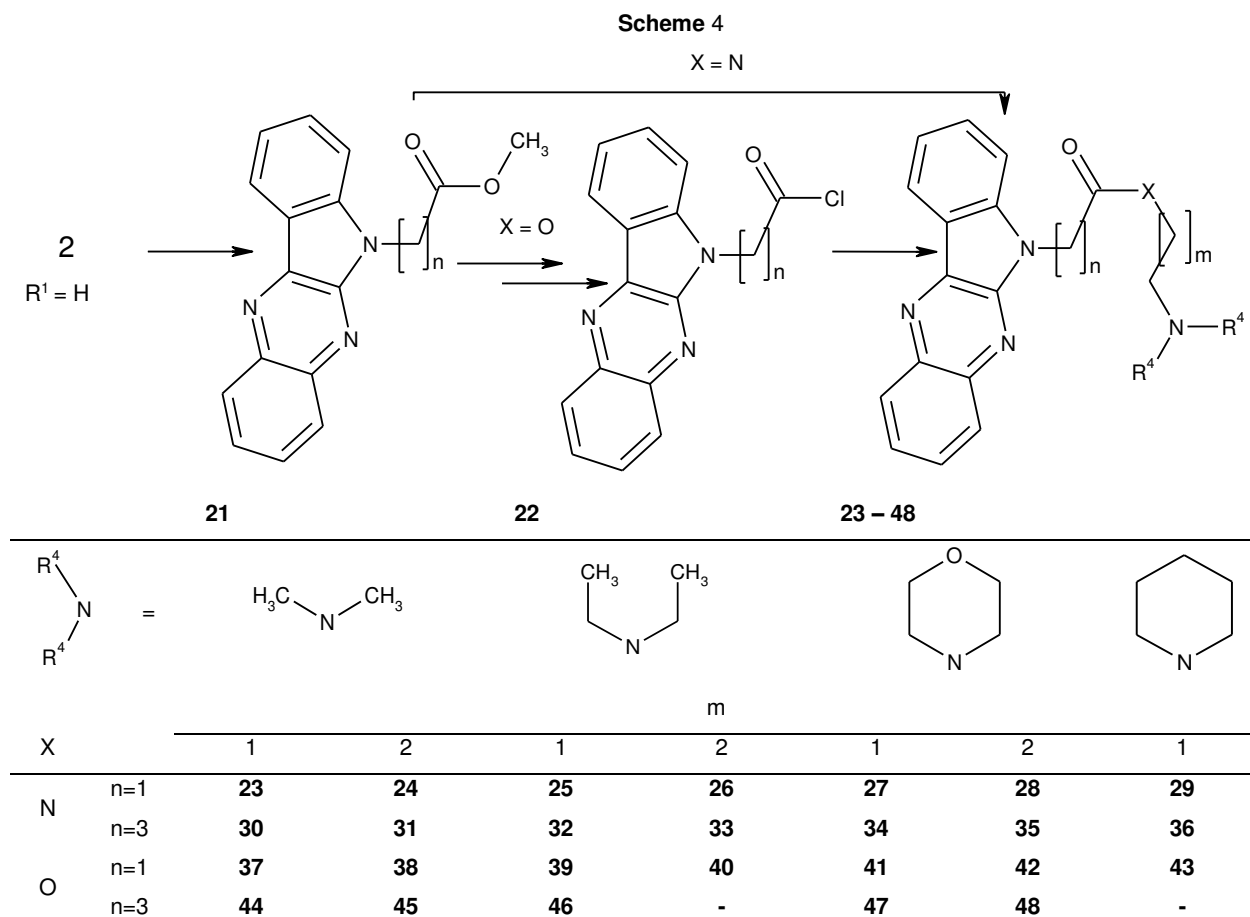


Alkylation of N,N-dialkylaminoalkylenediamines with **3** ($R^1 = \text{H}$) lead to aza-compounds **14 – 20** as it shown in **Scheme 3**:

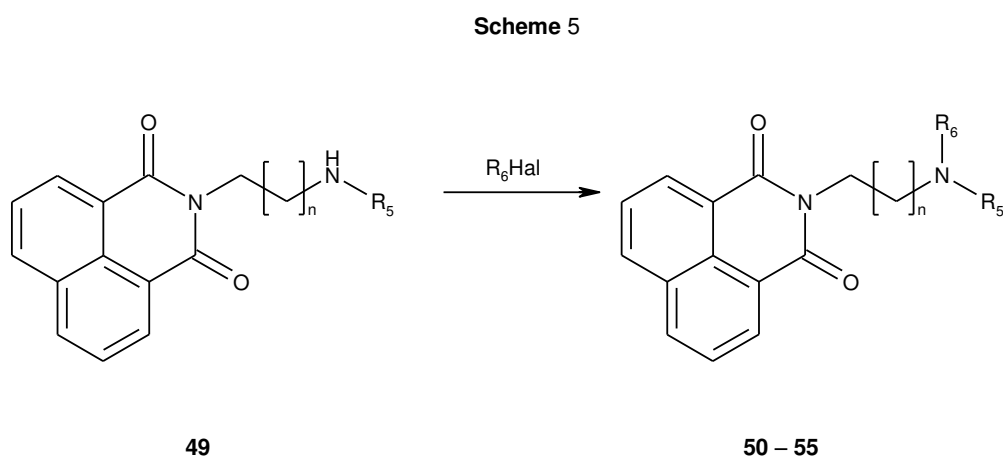


$\text{R}^3-\text{N}-\text{R}^3$	$n = 1$	$n = 2$
	14	15
	16	17
	18	19
	20	-

Treatment of indolo[2,3-b]quinoxaline **2** ($R^1 = H$) with bromoacetic acid methyl ester or 4-iodobutyric acid methyl ester in basic media leads to corresponding ω -(indolo[2,3-b]quinoxaliny)-carbonic acids methyl esters **21**. Esters after hydrolysis and drying were treated with thionyl chloride and obtained acyl chlorides **22** were converted into amides **23 – 36** ($X = N$)^[31] and esters **37 – 48** ($X = O$) during condensation with corresponding amines or alcohols (**Scheme 4**):

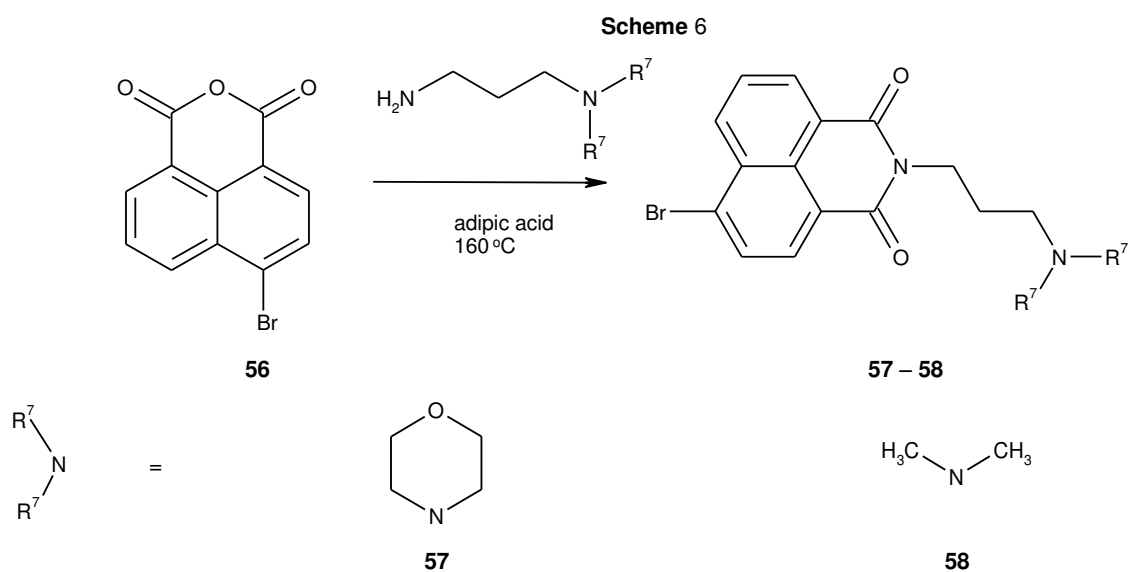


N,N-(dialkylamino)alkylnaphthalimides with different level of lipophilicity were obtained out of previously synthesized^[32] compounds by alkylation as shown on **Scheme 5**:



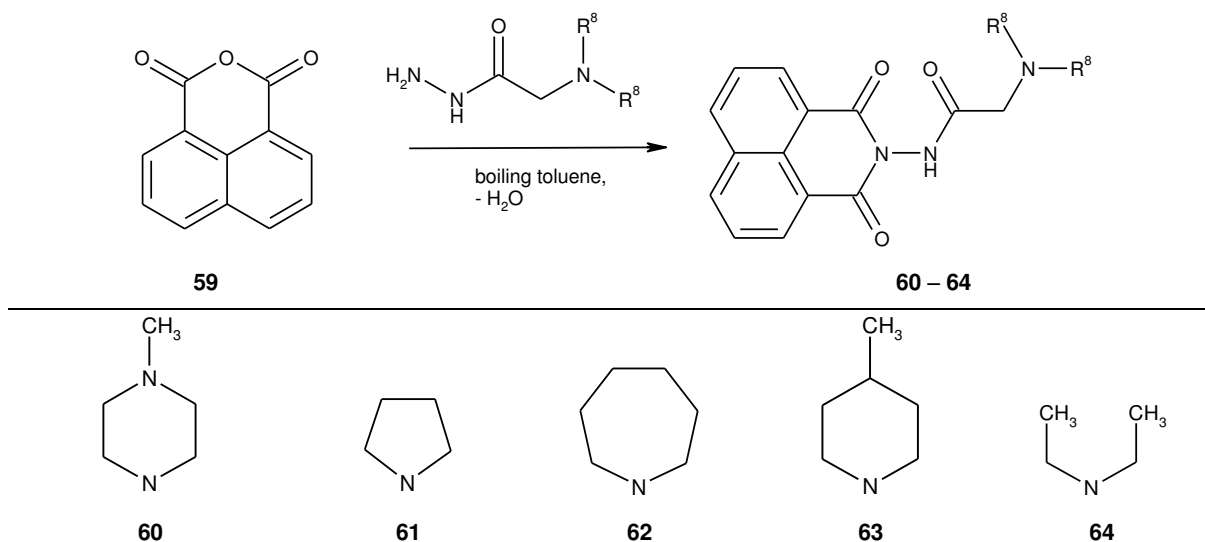
		n	
		1	2
		50	-
		51	-
R_6-N-R_5		-	52
		53	-
		54	-
		55	-

Compounds **57** – **58** were synthesized via condensation of 6-bromo-naphthalic anhydride **56** with corresponding diamines in hot adipic acid (**Scheme 6**).



N-aminonaphthalimide derivatives (**60** – **64**) were synthesized via condensation of naphthalic anhydride (**59**) with N,N-dialkylaminoacetylhydrazides in boiling toluene with azeotropic water elimination as it was shown in ^[33] (**Scheme 7**).

Scheme 7



The detailed description of synthesis of all compounds and corresponding analytical data and methods are provided in the Section 4 of the Supplementary Materials.

4. Biological tests

Compounds selected in virtual screening have been tested for their cytotoxicity via MTT assay and Real-time cell analysis. Antiviral activity against *Vaccinia virus* was determined using GFP expression quantitation and plaque forming units' assay. Interferon inducing capacity was assessed by decrease of a cytopathic effect caused by virus. Detailed description of experiments is provided in the Section 5 of the Supplementary materials.

15 out of 55 synthesized compounds were not soluble enough in 20% aqueous DMSO in order to complete sample preparation. The cytotoxicity of 40 compounds with respect to CV-1 cells was measured using MTT assay at samples concentration 0.1, 10, 100 μM . Cell viabilities observed for the most active compounds at concentration 10 μM are listed in Table 1. The values of CC_{50} (half-maximal cytotoxicity concentration) were evaluated for each sample at time point 24 h. Based on these data concentration 10 μM was chosen for screening of antiviral activity of the compounds because at this concentration we observed no or slight cytotoxicity for most of them.

Insertion of the DNA sequence encoding GFP into the thymidine kinase (TK) gene of *Vaccinia virus* significantly improves tracking of the virus without interfering with its ability to replicate. *Vaccinia virus* strain LIVP-GFP expressed GFP under the control of the early-late VACV VV7.5 promoter which resulted in efficient GFP expression during all stages of viral infection so that one can easily monitor the development of viral infection by measuring the level of GFP.

Compounds antiviral activity at 10 μM was evaluated in experiments with CV-1 cells infected with LIVP-GFP similarly to [35]. Data are listed in Table 2. CV-1 cells were treated with the compounds (10 μM) in duplicate. After 4 h of incubation the medium was removed and cells were infected by LIVP-GFP at a multiplicity of infection (MOI) of 0.01. LIVP-GFP-infected cells were incubated for additional 24 h prior to being processed for flow cytometry. The Relative expression of GFP is used for the primary assessment of the antiviral activity since it shows a decline in viral proteins formation (Table 2). The screening performed showed that five compounds (shown in Table 1) reduced GFP expression more than two times whereas other compounds lack GFP expression inhibition potency. These five were chosen for further testing.

Incubation of the cells with compounds **10**, **24**, **15**, and **19** at concentration 30 - 50 μM prior to infection results in 8 - 10 fold inhibition of GFP expression which reflects the strong antiviral activity of these compounds. Due to relatively high cytotoxicity of compound **17** (CC_{50} values 50) their antiviral activity was evaluated at concentrations not exceeding 10 μM : even at this relatively low concentration **17** twice reduce GFP expression level, thus showing rather pronounced antiviral activity.

We applied the plaque forming assay to analyze the effect of compounds **10**, **17**, **24**, **15**, **19** on viral infection development and infectious viral particles production in CV-1 cells. The virus titer was measured in the medium of infected cells pre-incubated for 4 h prior to infection with or without (control) above-mentioned compounds taken at different

concentration (Table 2). In parallel real-time monitoring of cell viability using xCelligence Real-Time Cell Analyzer was performed (see Figure A1 in Supplementary materials). CC_{50} values obtained for time-point 24 h for each concentration of the compounds used are shown in Table 1. Data obtained using xCelligence system are in a good agreement with the results of MTT test.

Table 1. CC_{50} and the antiviral activity measured by Flow cytometry analysis of GFP expression in the infected cells (Cells viability and Relative expression of GFP are given at concentration of 10 μ M). K- (negative control) – untreated, uninfected cells (cells autofluorescence). K+ (positive control) – untreated, infected cells. Information on all 40 compounds GFP expression test results is given in Table A6.

Compound ID	Cells viability, %	CC_{50} , μ M	Relative expression of GFP, %
K+			100
K-			8
10	79 \pm 10	\approx 100	23
17	74 \pm 8	\approx 50	28
24	84 \pm 16	100	43
15	88 \pm 9	\approx 50	45
19	79 \pm 6	\approx 50	48

Analysis of the antiviral activity showed that pre-incubation of CV-1 cells with the compounds resulted in the decrease of virus titer in cell medium by 05 – 1 lg(PFU/ml) and these results are consistent with the data obtained by flow cytometry. There are two lead compounds which exhibit antiviral activity in a concentration dependent manner, namely **10** ($\Delta_{\text{titre}} = 0.9$ lg PFU/ml) and **24** ($\Delta_{\text{titre}} = 0.8$ lg PFU/ml); for other tested compounds the differences in the viral titer were less pronounced ($\Delta_{\text{titre}} = 0.5 - 0.7$ lg PFU/ml). As for **17** ($\Delta_{\text{titre}} = 0.5 - 0.7$ lg PFU/ml) no dependence of the antiviral activity on the compound concentration was observed together with stimulation of cell proliferation. Compounds **15** and **19** inhibited viral infection only at the highest concentration used (30 μ M) by 0.5 – 0.6 lg PFU/ml and at this concentration 25% of CV-1 cells died.

Screening of antiviral activity shows that 2 out of 40 compounds tested, namely **10** and **24**, display prominent antiviral activity of appx. 10 folds decreasing infectious viral particles produced by infected cells. Compound **24** is characterized additionally by somewhat lower cytotoxicity in comparison with **10** (under similar conditions 90 and 75% of cell in population remained viable for **24** and **10**, respectively).

Antiviral activity of the studied compounds could be a result of either direct inhibition of viral infection by virus life cycle disruption or by inducing interferons (IFN) production by the cells. In order to analyze whether compounds work as IFN- α/β inducers we estimated the level of IFN in murine fibroblasts, infected with murine encephalomyocarditis virus (EMCV) after treatment with the selected compounds. Compounds under the study did not induce IFN- α/β production on a detectable level. Taking in account that efficacy of induction of IFN- α/β expression varied significantly in different cell lines we additionally tested induction of IFN- α in mouse spleen cells treated (stimulated) with the compounds **10**, **24**, **15**, and **19**. In these experiments no induction of IFN- α after the treatment mouse spleen cells with compounds was observed. Noteworthy, Cycloferon used as a positive control, stimulated IFN expression both in murine fibroblasts and in the mouse spleen cells.

Thus, two most promising compounds (Figure 4) **10** and **24** inhibit virus reproduction by at least 8 and 6 folds, respectively in considerably lower concentrations than their CC_{50} , which makes them eligible candidate for further antiviral research. Notice that their indolequinaxoline analogues from the training set antivirals occur more oftent (44 out of 62) among the most active compounds ($E_{\text{max}} \geq 50\%$).

The discovered hits were tested for DNA affinity (K_i) according to the procedure reported in^[15]. They display reasonable intercalating activity: $\lg(K_i) = 6.03$ and 5.20 for compounds **10** and **24**, respectively.

Table 2. Antiviral activity of potent compounds measured by classical plaque forming assay. n.d. – activity not determined. ¹⁾ Virus titer in the infected cell incubated in the presence of 0.002, 0.02 or 0.1% of DMSO in the cell medium was 3.1 \pm 0.3 PFU/ml, similar to K+.

Compound ID	C, μM	Viable cells, %	Virus titer, Ig(PFU/ml)
K ⁺ ¹⁾		100 \pm 0.1	3.3 \pm 0.1
10	1	107.8 \pm 0.1	3.1 \pm 0.1
	10	96.9 \pm 0.2	2.8 \pm 0.1
	50	75.6 \pm 0.1	2.4 \pm 0.1
24	1	103.6 \pm 0.2	3.4 \pm 0.1
	10	131.4 \pm 0.1	2.8 \pm 0.1
	50	90.0 \pm 0.1	2.5 \pm 0.0
15	1	113.7 \pm 0.1	3.5 \pm 0.1
	10	110.3 \pm 0.1	3.3 \pm 0.0
	30	75 \pm 0.0	2.7 \pm 0.1
19	1	107.2 \pm 0.1	3.6 \pm 0.2
	10	110.1 \pm 0.1	3.3 \pm 0.1
	30	75 \pm 0.0	2.6 \pm 0.1
17	1	97.1 \pm 0.1	2.8 \pm 0.0
	5	125.3 \pm 0.1	2.8 \pm 0.1
	10	125 \pm 0.0	2.6 \pm 0.1
	50	17.2 \pm 0.0	n.d.

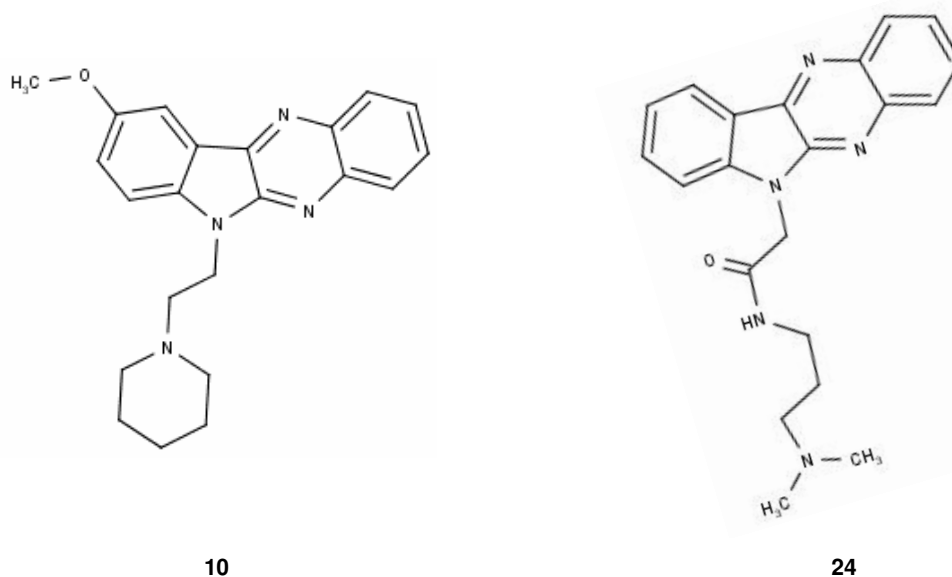


Figure 4. Prospective antiviral agent candidates

5. Conclusion

The multi-stage virtual screening workflow for computer-aided design of new broad spectrum antiviral agents acting as nuclear acids intercalators has been developed. Virtual screening involving structural filters, pharmacophore and QSAR models resulted in the hit list of 55 compounds structurally similar to those from the PCI set. These compounds have been synthesized and tested experimentally. 40 compounds from this hit list, soluble in 20% (DMSO) aqueous solution, were tested against Vaccinia virus - a double-strand DNA virus. Out of these, two molecules displayed high antiviral activity, reasonable

DNA affinity and low toxicity. Neither activity against a single-strand RNA virus nor interferon induction capacity have been detected for the studied molecules. The latter supports our hypothesis about intercalation mechanism of their antiviral activity.

Supplementary Material contains information about training set compounds structure and activity (Section 1), QSAR and pharmacophore model development and validation (Section 2), virtual screening performance (Section 3), synthesized compounds structure and purity (Section 4), biological test methods and activity (Section 4)

Acknowledgements. Kyrylo Klimenko thanks the French Embassy in Ukraine for the PhD fellowship. Elena Goncharova and Rinat Amirkhanov received financial support from ICBFM SB RAS.

6. References

- [1] S. Cleaveland, E. M. Fevre, M. Kaare, P. G. Coleman, Estimating human rabies mortality in the United Republic of Tanzania from dog bite injuries, *Bull World Health Organ*, 80 (2002) 304-310.
- [2] S. S. Shrestha, D. L. Swerdlow, R. H. Borse, V. S. Prabhu, L. Finelli, C. Y. Atkins, K. Owusu-Edusei, B. Bell, P. S. Mead, M. Biggerstaff, L. Brammer, H. Davidson, D. Jernigan, M. A. Jhung, L. A. Kamimoto, T. L. Merlin, M. Nowell, S. C. Redd, C. Reed, A. Schuchat, M. I. Meltzer, Estimating the burden of 2009 pandemic influenza A (H1N1) in the United States (April 2009-April 2010), *Clin Infect Dis.*, 52 (2011) Suppl. 1, S75-82.
- [3] M. P. Manns, T. von Hahn, Novel therapies for hepatitis C — one pill fits all?, *Nat Rev Drug Discov* 12 (2013), 595–610.
- [4] A.V. Ivachtchenko, Y.A. Ivanenkov, O.D. Mitkin, P.M. Yamanushkin, V.V. Bichko, I.A. Leneva, O.V. Borisova, A novel influenza virus neuraminidase inhibitor AV5027. *Antiviral Res* 100 (2013), 698–708.
- [5] S. Distinto, F. Esposito, J. Kirchmair, C. M. Cardia, E. Maccioni, S. Alcaro, L. Zinzula, E. Tramontano. Identification of HIV-1 Reverse Transcriptase Dual Inhibitors by a Combined Shape-, 2D-Fingerprint- and Pharmacophore-based Virtual Screening Approach. *Antiviral Res* 90 (2011), A31
- [6] F. Vigant, N. C. Santos, B. Lee, Broad-spectrum antivirals against viral fusion, *Nat Rev Drug Discov* 13 (2015), 426–437
- [7] E. De Clercq, Dancing with chemical formulae of antivirals: A panoramic view (Part 2), *Biochem Pharmacol*, 86 (2013) 1397-1410.
- [8] L. S. Lerman, Structural considerations in the interaction of DNA and acridines, *J Mol Biol* 3 (1961) 18-30.
- [9] L. Gao, M. Zu, S. Wu, A. L. Liu, G. H. Du, 3D QSAR and docking study of flavone derivatives as potent inhibitors of influenza H1N1 virus neuraminidase, *Bioorg. Med. Chem. Lett.*, 21 (2011) 5964-5970.
- [10] G. W. Rao, C. Wang, J. Wang, Z. G. Zhao, W. X. Hu, Synthesis, structure analysis, antitumor evaluation and 3D-QSAR studies of 3,6-disubstituted-dihydro-1,2,4,5-tetrazine derivatives, *Bioorg. Med. Chem. Lett.*, 23 (2013) 474-6480.
- [11] P. Zhan, C. Pannecouque, E. De Clercq, X. Liu, Anti-HIV Drug Discovery and Development: Current Innovations and Future Trends, *J Med Chem. Article ASAP*
- [12] T. M. Steindl, C. E. Crump, F. G. Hayden, T. Langer, Pharmacophore Modeling, Docking, and Principal Component Analysis Based Clustering: Combined Computer-Assisted Approaches To Identify New Inhibitors of the Human Rhinovirus Coat Protein, *J. Med. Chem.*, 48 (2005) 6250-6260
- [13] Z. Zhou, M. Khaliq, J.-E. Suk, C. Patkar, L. Li, R. J. Kuhn, C. B. Post, Antiviral Compounds Discovered by Virtual Screening of Small Molecule Libraries against Dengue Virus E Protein, *Chem. Biol.*, 3 (2008) 765–775
- [14] M. O. Shibinskaya, S. A. Lyakhov, A. V. Mazepa, S. A. Andronati, A. V. Turov, N. M. Zholobak, N. Y. Spivak, Synthesis, cytotoxicity, antiviral activity and interferon inducing ability of 6-(2-aminoethyl)-6H-indolo[2,3-b]quinoxalines, *Eur. J. Med. Chem.* 45 (2010) 1237-1243.
- [15] I. Antonini, P. Polucci, L.R. Kelland, E. Menta, N. Pescalli, S. Martelli, 2,3-Dihydro-1H,7H-pyrimido[5,6,1-de]acridine-1,3,7-trione Derivatives, a Class of Cytotoxic Agents Active on Multidrug-Resistant Cell Lines: Synthesis, Biological Evaluation, and Structure-Activity Relationships, *J. Med. Chem.*, 42 (1999) 2535 – 2541.
- [16] ChemAxon Standardizer. <http://www.ChemAxon.com/jchem/doc/user/standardizer.html> (accessed Feb. 2009)
- [17] D. Rognan, BioinfoDB: un inventaire de molécules commercialement disponibles à des fins de criblage biologique, *La Gazette du CINES* (2005) 1-4.
- [18] Calculator Plugins were used for structure property prediction and calculation, Marvin 6.1.4, 2013, ChemAxon (<http://www.ChemAxon.com>)
- [19] A. R. Leach, V. J. Gillet, R. A. Lewis, Taylor, R., Three-Dimensional Pharmacophore Methods in Drug Discovery, *J. Med. Chem.*, 53 (2010) 539-558.
- [20] E. Muratov, E. Varlamova, A. Artemenko, V. Kuz'min, P. Anfimov, V. Zarubaev, V. Saraev, O. Kiselev, QSAR Analysis of Anti-influenza (A/H1N1) Activity of Azolo-adamantanes, *Antiviral Res* 90 (2011), A74
- [21] P. Polishchuk, E. Muratov, A. G. Artemenko, O. Kolumbin, N. Muratov, V. E. Kuz'min, Application of Random Forest Approach to QSAR Prediction of Aquatic Toxicity, *J Chem Inf Model* 49 (2009), 2481–2488
- [22] G. Wolber, T. Langer, 3-d pharmacophores derived from protein-bound Ligands and their use as virtual screening filters, *J. Chem. Inf. Mod.*, 45 (2005) 160-169.
- [23] <http://zinc.docking.org/> v. 12

- [24] V. E. Kuz'min, E. N. Muratov, A. G. Artemenko, L. Gorb, M. Qasim, J. Leszczynski, The effects of characteristics of substituents on toxicity of the nitroaromatics: HiT QSAR study, *J Comput Aided Mol Des*, 22 (2008) 747-759.
- [25] A. G. Artemenko, E. N. Muratov, V. E. Kuz'min, N. A. Kovdienko, A. I. Hromov, V. A. Makarov, O. B. Riabova, P. Wutzler, M. Schmidtke, Identification of individual structural fragments of N,N'-(bis-5 nitropyrimidyl)dispirotriperazine derivatives for cytotoxicity and antiherpetic activity allows the prediction of new highly active compounds, *J. Antimicrob. Chemother.*, 60 (2007) 68-77.
- [26] L. Breiman, Random forests, *Machine Learning* 45 (2001) 5-32.
- [27] J. Jaworska, N. Nikolova-Jeliazkova, T. Aldenberg, QSAR applicability domain estimation by projection of the training set descriptor space: a review, *Altern. Lab. Anim.*, 33 (2005): 445-459.
- [28] K. Klimenko, V. Kuz'min, L. Ognichenko, L. Gorb, M. Shukla, N. Vinas, E. Perkins, P. Polishchuk, A. Artemenko, J. Leszczynski, Novel enhanced applications of QSPR Models: Temperature dependence of aqueous solubility, *J. Comput. Chem.*, 2016 DOI: 10.1002/jcc.24424
- [29] D.A. Filimonov, V. V. Poroikov, Prediction of biological activity spectra for organic compounds. *Russian Chemical Journal*, 50 (2006) 66-75.
- [30] <https://pubchem.ncbi.nlm.nih.gov/> (accessed February 2014)
- [31] M.O. Shibinskaya, E.A. Kovalenko, O.S. Karpenko, A.V. Mazepa, S.A. Lyakhov, S.A. Andronati, G.V. Antonovich, Z.M. Olevinsky, N.M. Zholobak, N.I. Spivak, E.V. Tretyakova, L.M. Shafran, M.Y. Zubritskiy, V.F. Galat, . Synthesis and affinity for DNA and interferon antiviral activity amides indolo [2,3-b] quinoxaline-6-yl-carboxylic acid, *Reports NAS Ukraine.*, 9 (2010) 125 – 131
- [32] O.S. Karpenko, I.V. Dorovskih, S.A. Lyakhov, S.A. Andronati, N.M. Zholobak, M.J. Spivak, Y.V. Nehoroshkova, L.M. Shafran, Aminoalkyl naphthalimides as interferon inducing and antiviral agents. *Synthesis and properties// Ukrainica Bioorganica Acta*, 6 (2008). 42 - 48.
- [33] O.S. Karpenko, S.A. Lyakhov, L.A. Litvinova, S.A. Andronati, M.N. Lebediuk, G.A. Horohorina, Synthesis and affinity of DNA N-[2-(dialkylamino)acetylaminonaphthalimides.], *Bulletin of the Odessa National University. Avg. Chemistry*. 10(2004), 176 - 183.
- [34] L. Rossetti, M. Franceschin, S. Schirripa, A. Bianco, G. Ortaggi, M. Savino, Selective interactions of perylene derivatives having different side chains with inter- and intramolecular G-quadruplex DNA structures. A correlation with telomerase inhibition, *Bioorg. & Med. Chem. Lett.*, 15 (2005) 413–420
- [35] K. Dower, K. H. Rubins, L. E. Hensley, J. H. Connor, Development of Vaccinia reporter viruses for rapid, high content analysis of viral function at all stages of gene expression, *Antiviral Res* 91(2011) 72-80
- [36] S. T. Nguyen, H. L. Nguyen, V. Q. Pham, G. T. Nguyen, C. D. Tran, N. K. Phan, P. V. Pham, Targeting specificity of dendritic cells on breast cancer stem cells: in vitro and in vivo evaluations, *Onco.Targets Ther.*, 8 (2015) 323-334.
- [37] A. Frentzen, Y. A. Yu, N. Chen, Q. Zhang, S. Weibel, V. Raab, A. A. Szalay, Anti-VEGF single-chain antibody GLAF-1 encoded by oncolytic vaccinia virus significantly enhances antitumor therapy, *PNAS*. 106 (2009)12915–20
- [38] W. Lowther, K. Lorick, S. D. Lawrence, W. S. Yeow, Expression of biologically active human interferon alpha 2 in Aloe vera, *Transgenic Res*. 21 (2012) 1349-1357.

Chemical Space Mapping and Structure–Activity Analysis of the ChEMBL Antiviral Compound Set

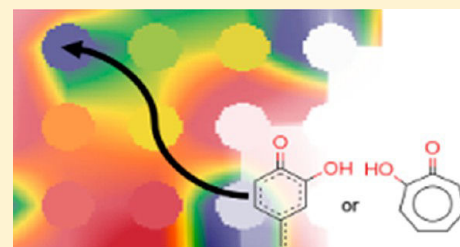
Kyrylo Klimentko,^{†,‡} Gilles Marcou,[†] Dragos Horvath,^{*,†} and Alexandre Varnek^{*,†}

[†]Laboratoire de Chimoinformatique, UMR 7140 CNRS/Université de Strasbourg, 1, rue Blaise Pascal, Strasbourg 67000, France

[‡]Department on Molecular Structure and Chemoinformatics, A.V. Bogatsky Physico-Chemical Institute of NAS of Ukraine, Lyustdorskaya doroga, 86, Odessa 65080, Ukraine

S Supporting Information

ABSTRACT: Curation, standardization and data fusion of the antiviral information present in the ChEMBL public database led to the definition of a robust data set, providing an association of antiviral compounds to seven broadly defined antiviral activity classes. Generative topographic mapping (GTM) subjected to evolutionary tuning was then used to produce maps of the antiviral chemical space, providing an optimal separation of compound families associated with the different antiviral classes. The ability to pinpoint the specific spots occupied (responsibility patterns) on a map by various classes of antiviral compounds opened the way for a GTM-supported search for privileged structural motifs, typical for each antiviral class. The privileged locations of antiviral classes were analyzed in order to highlight underlying privileged common structural motifs. Unlike in classical medicinal chemistry, where privileged structures are, almost always, predefined scaffolds, privileged structural motif detection based on GTM responsibility patterns has the decisive advantage of being able to automatically capture the nature (“resolution detail”—scaffold, detailed substructure, pharmacophore pattern, *etc.*) of the relevant structural motifs. Responsibility patterns were found to represent underlying structural motifs of various natures—from very fuzzy (groups of various “interchangeable” similar scaffolds), to the classical scenario in medicinal chemistry (underlying motif actually being the scaffold), to very precisely defined motifs (specifically substituted scaffolds).



1. INTRODUCTION

Viral epidemics are a present and serious threat to mankind,^{1,2} while antiviral compound research is one of the most challenging domains in drug discovery.³ There are several objective reasons for which modern research cannot promptly provide remedies against viral diseases, in particular the high mutation rate of viruses.⁴

During the last decades, significant research efforts have led to accumulation of relevant antiviral activity data. Advances in crystallography⁸ and extraction techniques⁹ made determination of viral proteins and nucleic acids structure possible. This information was crucial for target-based drug design^{20,21} leading to a breakthrough in drug discovery. In the era of “big data”, it is increasingly more difficult to exploit steadily accumulating experimental information, and to crystallize knowledge out of it. Electronic databases require *in silico* processing of chemical information, mining for recurrent patterns that may be useful knowledge for further rational drug development.

Even though commercial antiviral compounds databases exist,¹⁹ there is still considerable free access structure–activity data. Systematization and thoughtful description of this data is the main goal of current study. To this purpose, chemoinformatics provides a battery of tools for data curation and knowledge extraction. Here, we present an in-depth analysis of the structure–activity information relevant for antiviral compound research from the ChEMBL database,⁵ based on

chemical space mapping and structural pattern highlighting (detection of “privileged” key structural patterns and scaffolds encountered more often in antiviral compounds than in the rest of the ChEMBL collection, which here serves as a “reference” drug space).

Prior to this analysis, an extensive work of extraction and curation of relevant information from the heterogeneous, multisource activity data recorded in ChEMBL was followed by compound structure cleaning, standardization and duplicate removal. Eventually, we chose to conduct structure–activity analysis not with respect to each viral strain but to adopt a broader perspective based on virus classes. To this purpose, compounds that were reported active against virus strains of the seven best covered virus classes were grouped together into class-specific “positive” compound sets, whereas actives on less often encountered viral strains were labeled as “other antivirals”. In this way, homogeneous and large data sets were constructed on hand of multiple, sometimes small series of compounds tested against specific strains of specific viruses.

The analysis of the characteristics of the biologically relevant chemical space occupied by the considered classes of antivirals, and the detection of privileged structural motifs first requires the encoding of structures under the form of molecular

Received: April 7, 2016

Published: July 13, 2016

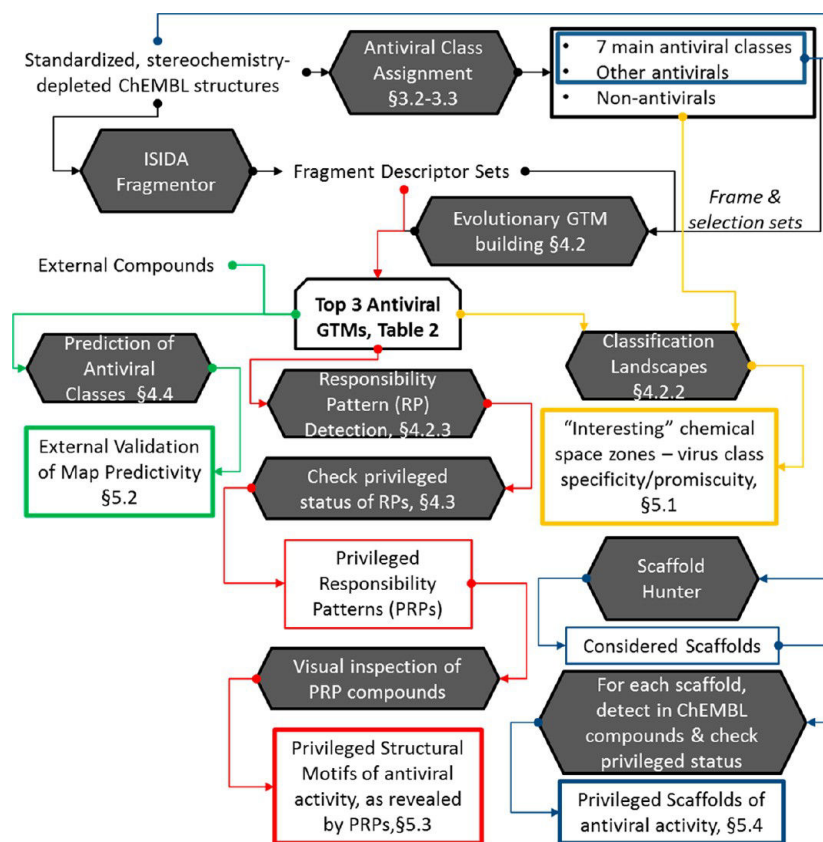


Figure 1. Workflow of the current study. Various connector colors denote the main tasks pursued in this work: in black, data curation, description and mapping; in yellow, separation of various compound classes into specific chemical space zones on the maps; in red, definition of responsibility patterns, highlighting of privileged responsibility patterns predominantly associated with an antiviral class and outlining the underlying structural motifs behind them; in blue, classical privileged scaffold analysis, for comparative purposes; in green, external validation, attempting prospective predictions of the antiviral class of external compounds.

descriptor vectors capturing relevant chemical information, i.e., the key structural features discriminating typical antivirals of a class from other antivirals and, respectively, from compounds without antiviral activity. However, it is not known, *a priori*, which molecular description scheme is best suited for such endeavor. Moreover, such molecular descriptor spaces tend to be high-dimensional, thus counterintuitive, and difficult to navigate or process. The evolutionary optimization procedure of dedicated generative topographic maps⁷ (GTM), designed for the purpose of simultaneous selection and dimensionality reduction of possible descriptor spaces, was used here in order to develop chemical space maps of maximal relevance for the specific, herein-considered problem of defining the specific traits for antiviral compound groups.

GTMs produce two-dimensional, readable depictions of the compound collections. They provide a nonlinear dimensionality reduction of the initial descriptor space, into a two-dimensional square grid of “nodes”, onto which each compound will be projected. The size of this grid, as well as other technical parameters controlling map and—paramount—the molecular descriptors chosen to encode structural information were considered as degrees of freedom in an already published⁴⁵ evolutionary map “growing” procedure. The output of the procedure is a population of near-optimal maps, out of the pool of all the possible maps that could have been built with the initially proposed sets of ISIDA molecular descriptor spaces.^{46–48} The objective quality criterion for resulting maps reflects the ability of a resulting map to

“separate” the defined antiviral classes, i.e., to project compounds belonging to each antiviral class into specific, dedicated areas of the 2D-grid. This is the same line of reasoning defended in a recent publication⁴⁵ aimed at optimizing “universal” GTMs of broad polypharmacological competence—ability to separate actives from inactives for a maximum of completely independent biological properties. Maps produced as a part of that study were actually proven to be able, as an external test, correctly regroup the herein studied antiviral compounds by virus class—although antiviral activities were not used to select them. However, that work did not pursue the in-depth analysis of trends revealed by those maps. Now, the same evolutionary procedure was specifically applied to antiviral compounds, with a problem-specific “antiviral” optimality criterion defined in terms of balanced accuracy of separation of above-mentioned virus-group specific compounds. Therefore, the evolved “antiviral” maps, expected to show improved separation power compared to previously published “universal” maps, are better suited to highlight common structural motifs corresponding to specific map zones with high antiviral compounds densities. Mapping highlights both specific and overlapping chemical space zones, for specified classes, and various classification landscapes—highlighting one antiviral class by contrast to all the other antivirals, or one antiviral class versus another one, or all antivirals, on one hand, versus the remainder of nonantivirals from the ChEMBL database on the other, *etc.* The zone-specific dominance of a class over the other, as encoded by the prevalence of class-

assigned color codes, and/or the “promiscuous” zones where the classes are not well separated, can be read from the maps and used in prospective predictions of antiviral properties of novel compounds. A proof-of-concept prediction exercise has been carried out.

Common structural motifs are the underlying reasons for which compounds are being projected onto a same GTM zone, defined by the so-called responsibility vectors, which reflect the probability of a given compound to “reside”—in a fuzzy-logics acceptance of this term—on each of the nodes defining the 2D grid. Structurally similar compounds will be preferentially mapped with similar responsibility values onto the same nodes.

Note that the quest for “privileged” structural patterns^{49,50} is a hallmark of recent, rational drug design techniques. A structural motif is considered class-privileged if its occurrence rate in active compounds of a therapeutical class (actives containing the motif/all actives of the class) is much larger than the “default” occurrence of the motif (any compound containing the feature/all representatives of the reference compound set). It is fairly easy to decide whether a predefined structural motif is “privileged” or not, according to the definition above. The real challenge is how to “guess” what precise structural motifs—or even what class of motifs (scaffolds, generic substructures, subgraphs, pharmacophore, molecular field patterns)—one should actually submit to the privileged status check in order to find the relevant ones, of the infinity of imaginable ones. There is a tendency in medicinal chemistry to focus on scaffolds—a potentially biased view, because the scaffold alone may be only partially responsible for activity or, on the contrary, different scaffolds might be biososteric.⁶ By contrast, the herein introduced GTM-driven responsibility pattern analysis is able to suggest structural motifs of maximal relevance: it is sufficient to highlight privileged responsibility patterns (PRPs) that are preferentially seen in (“privileged” by) antiviral compounds of any given class, and then highlight—by simple visual inspection—the underlying privileged structural motifs (PSM), shared by most of the same-pattern molecules. PSMs behind the PRPs might, but do not need to, coincide with scaffold definitions: they may be more specific (pattern = given scaffold plus a specific substitution pattern) or, on the contrary, fuzzier (regrouping similar scaffolds, or referring broadly to a common pharmacophore pattern that may be ported by various scaffolds). In complement to a classical scaffold-oriented analysis, the highlighted privileged responsibility patterns are an excellent example of how chemical space mapping using GTM allows rationalizing structure–activity information.

A workflow of visualization and analysis of chemical space of antiviral compounds is given in Figure 1.

2. METHODS

2.1. Data Preparation. The ChEMBL database⁵ (as of November 2014) was picked as one example of a high-quality, publicly available bioactivity database, and used as the primary data source in the current study.

2.1.1. ChEMBL Compound Structure Extraction, Standardization and Description. All organic structures from the ChEMBL database were extracted and standardized according to the in-house rules implemented on our virtual screening web server, powered by the ChemAxon⁵⁸ toolkit (removal of heavy-metal-containing species, of enormous molecules at >100 heavy atoms, salt removal, conversion into the predicted, most stable tautomer form, representation of N oxides with split formal

charges, conversion to the “basic” aromatic forms of five- and six-membered aromatic rings, *etc.*). Because herein used molecular descriptors—fragment counts—are not capturing stereochemical information, a list of standardized, unique (~1.2 million) stereochemistry-depleted SMILES⁵⁹ strings (in which any special characters denoting chirality or cis–trans isomerism were removed) was set to represent the reference chemical space for this study. Note that some of the herein stored unique SMILES strings may be linked to several ChEMBL ID values, which may correspond to different stereoisomers mapped onto a common stereochemistry-depleted SMILES, to different formulations (counterions in salts) accompanying the same active principle, or simply to genuine duplicates of a same structure under different ChEMBL IDs in the original database.

As already mentioned in previous work,⁴⁵ the 38 different ISIDA fragmentation schemes considered to be reasonable initial choices of chemical spaces were generated for the 1.2 million unique structures.

2.1.2. From Brute ChEMBL Activity Data to Antiviral Activity Label Assignment. The first step was querying ChEMBL for antiviral activity-related information via the public web interface using “antiviral” as a query word. The query result was downloaded as a CSV file containing 52 columns with data description (e.g., target type, compound’s ID, compound’s canonical SMILES, *etc.*) and 114 324 rows of data entries, pertaining to 35 547 compounds, in the sense of distinct ChEMBL IDs. Activity data extraction and curation was performed with the KNIME¹⁰ software. This program operates as sequences of nodes in which data manipulation takes place. KNIME was used to implement filtering rules to data table from CSV file, which contained ChEMBL antiviral activity data. Rules were created in order to filter out inconsistent data, removing entries if:

- Data validity comment contains the following keywords: “Outside typical range”, “Potential missing data”, “Non standard unit for type”. 33 370 compounds (distinct ChEMBL IDs) remained.
- Activity comment does not report “active”, further decreasing compound number to 32 431.
- Potential duplicate column signals redundant data, 32 420 compounds remained.
- Assay type contains “ADME” (thus reporting pharmacokinetics data on the host, by contrast to functional¹¹ and binding¹² assays, which were kept, 32 373 molecules).
- Assay CRC description was not set to “Scientific Literature”, 32348 compounds remained.
- Standard type, the nature of the reported activity score points to rarely occurring activity measurements such as replication efficiency,¹³ or activity parameters that were not actually related to antiviral activity. More detailed information concerning this field is given in Supporting Information Table S1, together with the specific thresholds defining “active” status with respect to the standard value field, which must be interpreted in relation to the standard type and standard units fields. A total of 24 629 molecules were still selected after this last stage.

The next step was definition of the major antiviral classes. It required an analysis of all the different antiviral assay protocols reported in the filtered entries, in order to define a clear grouping criterion for associating activities from particular

assays with antiviral classes. In such a way, each compound found active in an assay becomes a “positive” representative of the class to which the assay was assigned. Because the number of viral protein targets is very large, and not all antiviral compounds have a strictly defined mechanism of action, pathogen type was chosen to be the core of antiviral classification. Information on virus types is given in the assay organism column. Entries with the assay organism field matching one of the seven text-mining queries below (an asterisk matching any pre- or postfix characters) were assigned into corresponding antiviral classes (details given in Table 1).

As a consequence, ChEMBL compound IDs could be linked to the seven specific antiviral classes, plus the “other antiviral” class containing antiviral agents against any other viral strains. If a specific ChEMBL ID was associated with one (or more) of the seven major viral classes, it was labeled as “positive” with respect to the class(es). The “negative” status with respect to a class was assigned to a given ChEMBL ID if it represents a positive associated with another class or it has not been recognized as positive at all. Compounds labeled “other antivirals” systematically appear among the “negatives” associated with the seven main classes.

KNIME-driven activity data curation thus yielded 49 191 reliable data entries, comprising information on 24 629 unique ChEMBL smiles strings. These correspond to 24 633 different compounds (in the sense of distinct ChEMBL compound ID values), published in 1982 papers. Four compounds^{14,15} were erroneously duplicated in ChEMBL, and given different ChEMBL ID for same structures.

Note that experiments described in articles accounted for by ChEMBL were carried out by different research teams following different experimental workflows. The lack of clear ontology and homogeneity of antiviral data made grouping of antivirals a challenge. For this reason, but also because the latter strategy leads to larger data sets (providing much-needed statistical robustness for further analysis), a higher-level merging of ChEMBL sets—by viral class membership and irrespective of specific assay conditions—has been preferred in this work.

ICTV’s “Viral Taxonomy: 9th report”¹⁶ was used to choose the classification criterion among existing taxonomic ranks from *Order* to *Species*. Viruses are relatively simple organisms and even slight changes in virion structure can lead to dramatic shift in pathogenicity; therefore, it is hard to define the clear difference between them. The most coherent taxonomic rank for virus is *Family* because it is based on criteria, such as

- Genome nature (e.g., dsDNA, ssRNA(-))
- Envelope (presence or absence)
- Morphology (i.e., virion form)
- Virion form (e.g., bullet-shape)
- Genome configuration
- Genome size
- Type of host organism

However, *Family* is insufficient for grouping because none of the mentioned criteria takes into account protein composition of the virus, making structure–activity determination more complicated. Therefore, lower levels of hierarchy, namely *Genus* and *Species*, were examined. Further consideration revealed that both of these taxonomic ranks take virion protein composition into account but *Genus* allows grouping more viruses of similar origin into one category (e.g., HSV-1 and HSV-2 have similar protein composition but they are assigned to different *Species*

because HSV-1 causes mostly sore colds and HSV-2 causes mostly genital herpes).³⁶ Thus, virus *Genus* was chosen as a classification criterion, resulting in the definition of 7 major activity classes. *Genera* of major classes include only mammalian viruses.

The compiled antiviral classes (Table 1) comprise the most dangerous and widespread viral pathogens.

Table 1. Text Queries Used To Classify ChEMBL Compound–Activity Records into Antiviral Classes, On Hand of the Assay Organism Entry^a

antiviral class	assay organism matches:	hit count
enterovirus (<i>Ent</i>)	“*human rhinovirus*” OR “*human enterovirus*”	424
hepacivirus (<i>Hep</i>)	“*hepatitis C virus*”	5320
influenza A (<i>Inf</i>)	“*H2N2 subtype*” OR “*influenza A virus*”	638
lentivirus (<i>Len</i>)	“*HIV*” OR “*human immunodeficiency virus*”	8854
Orthohepadnavirus (<i>Ort</i>)	“*HBV genotype D*” OR “*hepatitis B virus*”	700
pestivirus (<i>Pes</i>)	“*bovine viral diarrhea virus 1*”	412
simplexvirus (<i>Sim</i>)	“*human herpesvirus 1*” OR “*human herpesvirus 2*” OR “*Hsv-2*” OR “herpes simplex virus (type 1/strain F)*”	790
other antivirals	entries not matching any of the above	7897
total antivirals	sum of above, compounds present in several classes counted only once	24629

^aThe total compound count on the last line is lower than the sum of listed class members, because some compounds may be members of several classes.

2.1.3. Linking Structure to Activity Class. Curation proceeded according to two separate workflows: of chemical structures, and of activity information, respectively. Eventually, structural data has been related to activity information. Each unique standardized and stereochemistry-depleted compound structure (SMILES string) corresponds to the (one or several) ChEMBL IDs. For each of the 1.2 million standardized compounds, the ChEMBL IDs were searched within the listed “positives” and “negatives” associated with each of the seven virus classes. If none of the ChEMBL IDs of a standard compound is present in that list, that compound was classified as “outside” the antiviral chemical space, and labeled “0”. If at least one of the ChEMBL IDs is present among the entries of a class, then the compound was classified as positive with respect to that class. Negatives of each class are, by contrast, all the positives of other classes—except for the “promiscuous” that are actually listed as positive for the current class too—and the “other antivirals”. Eventually, an antiviral profile text file *profile_antivir.dat* (see the Supporting Information) has been compiled for the entire 1.2 million ChEMBL collection of standardized, stereochemistry-depleted SMILES strings. It is a seven-column file, each line corresponding to a structure *M*, in the order listed in *StdChEMBL.smi_chid*, the Supporting Information file listing standard SMILES strings associated with their ChEMBL ID code(s)—concatenated by the “+” sign if more than one ChEMBL ID corresponds to a SMILES string. Each column corresponds to an antiviral class *C*, in alphabetical order as given in Table 1. Status labels in this matrix, *Stat(M,C)* may be “2” if *M* is a positive of class *C*, “1” if *M* is a negative of *C*, or “0” if *M* is a structure outside of the antiviral chemical space—in this case, $Stat(M,C) = 0 \forall C$.

It is important to note upfront that a categorization as “positive” is a clear statement that this substance has an antiviral effect on at least one member of the given viral class. “Negative” means, in most cases “unknown activity” for that class, as in case when the compound was not tested against a particular group. The “negatives” are to be used as examples of compounds associated with different viral classes, in an attempt to learn what chemical features differentiate the drug candidates against one class of viruses from those associated with another. Although “positive” is synonymous to “active”, it is not reasonable to interpret “negative” as “inactive”, especially in this context where labels do not refer to a specific viral strain, but to a whole class. Formally, a compound could be declared “inactive” against a group only if it would be tested and found inactive against each virus of that group—an impossible endeavor. Also, note that ChEMBL compounds that were never reported to participate in any antiviral tests, and herein considered “outside” of antiviral chemical space, are distinguished from “negatives”, even though they are also likely to be inactive. The difference is that a “negative” has the peculiarity to be considered, by at least one group of scientists, as relevant enough in order to deserve being screened against at least one viral strain. “Negative” should be interpreted as “presumably different” from the effective antivirals of given virus class, all while being interesting antiviral compounds, targeting other viruses. Therefore, if, for example, a standardized, stereochemistry-depleted structure is associated with two ChEMBL IDs, one of which (ID1) is reported as “positive” against viral group 1, whereas ID2 is given as “positive” against another group 2, careful investigation is needed. Apparently, this is a paradoxical situation, because ID1 as “positive” for group 1, and not encountered in any measures run against group 2, would be by definition be assigned as a “negative” of group 2, and *vice versa*. Or, both IDs refer to a common standardized structure, i.e., to a common point in the vector space of stereochemistry-ignorant molecular descriptors. If ID1 and ID2 represent a case of genuine compound deduplication (compound has no stereoisomers, but perhaps comes in different formulations, as salts with different counterions, etc.), it is safe to assume that, although some authors have used the substance ID1 to report their test on class 1, the test on class 2 running under ID2 concerned exactly the same compound. It is thus safe to decide that the compound is “positive” with respect to both classes, the apparent problem being due to ChEMBL, having referred to it by different IDs. If, however, the compound has stereoisomers, one should check if the two different testing campaigns did actually involve the same isomer. However, the goal of the current paper is not to investigate aspects of stereochemistry, because it is out of used descriptors applicability, but the robust structural features associated with antiviral activity. After structure standardization and generation of stereochemistry-depleted unique SMILES codes—only 2.5% of compounds were associated with more than one compound ChEMBL IDs, i.e., represent potential stereochemistry-related issues. This number is conveniently small to justify employment of 2D molecular descriptors in this work.

The herein employed classification scheme—“positives” vs “negatives” for each virus class, plus ChEMBL compounds “outside” antiviral chemical space—is thus not an experimental activity profile matrix, in which each compound \times target table cell represents a measured activity value. Such a matrix should have had specific virus strains listed as targets, and would have been extremely sparse, thus virtually useless for robust analysis

of structure–activity trends. The empirical classification advocated here is not an absolute record of antiviral activity facts, but a snapshot, prone to further evolution, of what is known for sure to work and what medicinal chemists would expect to work against virus groups. This coarse view has the merit of robustness, and if chemoinformatics may prove that the above-mentioned chemical subspaces are well distinct, i.e., present specific, privileged structural patterns, these latter will represent a way to sketch, in broad lines, the *status quo* of present-day antiviral research.

2.2. Scaffold Detection. Scaffold Hunter software¹⁷ was used to define a compound’s chemical class by determining the most common substructures within the major activity classes—except the set of lentivirus-positives, which is too large and too diverse for the present purpose. This software organizes scaffolds in a tree-like hierarchy based on the inclusion relation, enabling navigation in the associated chemical space in an intuitive way. A standard¹⁸ tree-forming procedure was carried out in this study.

2.3. GTM Construction. Optimally discriminating GTMs were built following the same evolutionary strategy⁴⁵ used to generate “universal” maps of maximal generality for the entire drug space.

2.3.1. GTM Algorithm: An Intuitive Outline. Generative topographic mapping (GTM)^{51–54} is a dimensionality reduction algorithm that fits a “rubber sheet” (a bidimensional manifold) in the initial vector space defined by molecular descriptors, in which every molecule is located at its specific point. The algorithm “distorts”—within the allowed limits provided by control parameters—the rubber sheet in such a way as to make it touch, or approach, as many of the molecular points, which are then englobed in the manifold or projected onto its nearest point. The rubber sheet is then “straightened out” onto a bidimensional square grid, and the projections of the molecules are fuzzily associated with the nearest grid points. A probability matrix $R(M,K)$ of molecule M residing onto the grid node K (technically “responsibility of node K for molecule M ”) is calculated. A molecule may be a “full-time resident” of a single node K : $R(M,K) = 1.0$; $R(M,K') = 0.0 \forall K' \neq K$, (technically, it has a “one-node” responsibility distribution) or it may be distributed over several nodes (“many-node” distribution). In either case $\sum_K R(M,K) = 1.0$, the cumulated probability to see a compound anywhere on the map is one.

2.3.2. Cumulated Responsibility (Distribution Density) and Classification Landscapes. A compound set S can be characterized on a GTM by the cumulated responsibility vectors of all its members: $\rho(S,K) = \sum_{M \in S} R(M,K)$. The notation ρ for cumulated responsibility was chosen on purpose, to highlight that this magnitude is nothing else but the node-bound density of distribution of the compound set, representing the fuzzy count of the numbers of members of set S residing in each node of the GTM. In a “null model” GTM providing no meaningful mapping—all compounds of the set being equally distributed over all the nodes—the baseline density $\rho^0(S)$ at each node would equal the ratio between the number of compounds of S and the total number of nodes of the map. Thus, for two unbalanced sets, the densities of the larger one will mechanically be larger than the ones of the small one—instead of the comparative mapping of the brute ρ scores, it is advisable to focus on normalized densities $\rho^*(S) = \rho(S,K)/\rho^0(S)$. A node K is said to be predominantly populated by compound set S if $\rho^*(S,K) > \rho^*(s,K)$, for any other benchmarked set s . In particular, if S and

s are taken to be the sets of positives “2” and negatives “1” associated with an antiviral class, the for each node a fuzzy, mean classification score can be obtained as

$$\bar{C}(K) = \frac{2 \times \rho^*(2, K) + \rho^*(1, K)}{\rho^*(2, K) + \rho^*(1, K)} \quad (1)$$

where \bar{C} will be closer to 2 in predominantly “positive” nodes, and closer to 1 in the others. Nodes may hence be colored by relative predominance of positives versus negatives, and in the present work a five-color spectrum-based representation was used to highlight “negative” zones ($\bar{C} < 1.4$) in red, “slightly negative” ($1.4 \leq \bar{C} < 1.5$) in orange, “slightly positive” ($1.5 \leq \bar{C} < 1.6$) in yellow, “positive” ($1.6 \leq \bar{C} < 1.7$) in green, and “strongly positive” ($1.7 \leq \bar{C}$) in blue. A finer chromatic resolution was used for the “positive” map areas—this is easily tunable.

Given the fuzzy nature of responsibility vectors, ρ values for any given node may be arbitrarily low, but never zero, hence eq 1 is applicable to every node of the map, which can be colored by its mean, fuzzy propensity to host molecules of a given class. However, for nodes on which no molecules do actually reside—where the ρ values represent practically meaningless distribution tails—the calculated \bar{C} values make no chemical sense, and should not be represented. In this work, it was chosen to modulate color intensity (the alpha channel) by the total compound density $\rho(1,K) + \rho(2,K)$ at a node, so that color rendering can be tuned from completely transparent (if total density is below a minimal threshold) to full saturation (if density exceeds a maximal threshold), with range-wise interpolation in between. Thresholds in this work were chosen with respect to the total size of the mapped compound sets. On maps featuring the entire 1.2 million ChEMBL set, nodes with total densities below 1 are fully invisible, whereas more than 10 000 compounds are needed to render a node at full color saturation. On maps representing only the antiviral set, or specific antiviral class subsets, minimal and maximal density thresholds were 0.1 and 10.0, respectively.

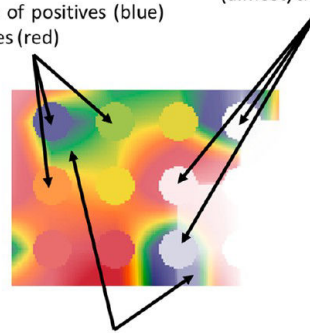
Eventually, note that on a GTM, unlike in a Kohonen map, continuous property “landscapes” over the 2D space covered by the grid of nodes can be interpolated. In this work, nodes are represented by circles of homogeneous color and density, reflecting the property and density values at the node, whereas color and density in the internodal continuum are obtained by polynomial interpolation. A zoom-in on a typical classification landscape is given in Figure 2.

In the present work, various classification landscapes were generated. First, antiviral class-related landscapes distinguishing between positives (“2”) and negatives (“1”) for each of the seven antiviral classes will also be used for predicting the putative class membership of novel compounds, thus providing a mechanism for external model validation. However, formal class labels 1 and 2 can be reassigned in order to monitor the generic chemical space occupied by the entire antiviral set (now collectively assigned to class “2”) by contrast to the rest of the ChEMBL database (outside of antiviral chemical space, now class “1”). Single class plots may also be realized, where the only variable is density.

2.3.3. Responsibility Patterns: Definition. Two molecules with identical (or similar) responsibility vectors are undistinguishable (or close) on the built map and should, therefore, have similar properties; otherwise, this is a low quality map, not complying with the principle of neighborhood behavior.^{55,56}

Nodes – circles of color reflecting relative predominance of positives (blue) vs. negatives (red)

Empty (low density) nodes are rendered (almost) transparent



In the continuum between nodes, both color and transparency are interpolated on the basis of the four closest nodes

Figure 2. Zoom-in on a GTM-based classification landscape.

Because $R(M,K)$ is a real-value matrix, the chance to find two molecules with strictly identical responsibility vectors is very low. It is thus convenient to introduce discretized herein termed “responsibility pattern” vector RP , in replacing actual R values by standardized responsibility level indices. If the responsibility of M for node K is below what is empirically considered “below the minimally relevant threshold”—empirically established at 1%, the corresponding integer responsibility level is set to zero. Beyond this threshold of 0.01, any additional 0.1 units of responsibility contribute an increment of +1 to the RP value, i.e., $RP(M,K) = 1$ if $0.01 \leq R(M,K) < 0.11$, $RP(M,K) = 2$ if $0.11 \leq R(M,K) < 0.21$, etc. Formally, one may therefore define:

$$RP(M, K) = [10 \times R(M, K) + 0.9] \quad (2)$$

where the $[..]$ operator means truncation. Molecules with the same RP vector are considered to be members of a same responsibility cluster—unavoidable binning artifacts notwithstanding—the reason for which RP is referred to as the “responsibility pattern” of a given molecule, and hence of its associated cluster. This procedure is nothing but cell-based clustering⁵⁷ in responsibility vector space, at 10-fold split of each descriptor component range.

2.3.4. GTM Fitting. The herein built maps focus on the chemical space of antiviral compounds, which in this work are at the basis of both frame and selection sets.

The previously mentioned⁴⁵ 38 different ISIDA fragmentation schemes provided descriptor choices for the Darwinian selection procedure.

Frame sets are compound collections used to generate the GTM manifold. Thus, these need to be chosen such as to span the entire relevant zone of the chemical space, therefore providing points of support for a robust fitting of the manifold. Here, subsets of the antiviral set, of various sizes, were taken, independent of compound assignment to antiviral classes (frame sets do not convey any activity-related information, because manifold construction is completely unsupervised). As automatic frame set selection is also a degree of freedom of the evolutionary map building procedure, three different frame set choices were considered in this case: (1) the entire antiviral set, (2) half of the antiviral set (every second entry) and (3) a quarter of the viral set (one compound out of four).

During the evolutionary procedure, the fitness score used for map selection was based on the 3-fold cross-validated

Table 2. Top Antiviral Maps Emerged from Darwinian Optimization

map	descriptors	size ^a	cross-validated balanced accuracies/antiviral class						
			<i>Ent</i>	<i>Hep</i>	<i>Inf</i>	<i>Len</i>	<i>Ort</i>	<i>Pes</i>	<i>Sim</i>
1	IIAB-1-3 ^b	28 × 28	0.90	0.83	0.82	0.79	0.83	0.86	0.82
2	IIRA-P-1-5 ^c	34 × 34	0.89	0.84	0.81	0.77	0.83	0.85	0.80
3	IIAB-FF-1-2 ^d	42 × 42	0.91	0.83	0.79	0.77	0.80	0.86	0.83

^aMap sizes are reported as numbers of nodes per line of the square grid. Please see Supporting Information Table S2 for the other technical parameters used to build them. ^bWinning ISIDA fragmentation scheme of circular atom and bond fragments of size (topological radii) between 1 and 3. For more details, see ISIDA fragmentation scheme nomenclature.⁴⁶ ^cWinning ISIDA fragmentation scheme of circular pair counts of sizes 1 to 5. For more details, see ISIDA fragmentation scheme nomenclature.⁴⁶ ^dWinning ISIDA fragmentation scheme of circular atom and bond fragments colored by the CVFF force field type. For more details, see ISIDA fragmentation scheme nomenclature.⁴⁶

propensities to discriminate positives of each of the seven antiviral classes from its negatives (which make up the rest of the antiviral set). In other words, the seven different “selection sets” were represented by the same antiviral compound set, but in association to the seven different status labels. As the selection is driven by success in seven different classification tasks, the overall success score (map fitness score) is calculated on the basis of individual balanced accuracies for each task, as their mean value penalized by their standard deviations (details in previous publication).

Out of the top maps emerging from the Darwinian evolution simulation, the best three based on distinct descriptor choices were selected. Integer responsibility patterns for each antiviral compound were determined on each map according to eq 2. Likewise, ChEMBL molecules outside the antiviral space (labeled class “0”) were retrospectively mapped, and their responsibility patterns extracted.

2.4. Privileged Pattern Detection. Compounds featuring a common responsibility pattern (*RP*) on a GTM, as well as compounds sharing a same scaffold, can be regarded as a “cluster”. If such “cluster” contains a significantly high percentage of molecules associated with a given activity class (with respect to the entire library), then the pattern defining the cluster (*RP*, scaffold) is privileged^{49,50} with respect to that activity class. Let $f_{\text{act}}(RP)$ represent the fraction of “active” compounds matching a given pattern *RP*, where “active” should here be understood in the broad sense of molecule having a desired property, belonging to a given therapeutic class, having a special status. By contrast, let $f_{\text{def}}(RP)$ represent the default fraction of molecules, out of the entire collection under study and related to the pattern. A privileged pattern *RP* associated primarily with the “actives” will have $f_{\text{act}}(RP)/f_{\text{def}}(RP) \gg 1$.

Here, two distinct types of “privilege” will be defined. On the one hand, one may check whether a pattern is seen more often within all antiviral compounds (positives of the seven classes, plus other antivirals), with respect to the entire ChEMBL database. This antiviral specificity score (*Asp*) can be thus defined as

$$Asp(RP) = \frac{f_{\text{antiviral}}(RP)}{f_{\text{ChEMBL}}(RP)} \quad (3)$$

where $f_{\text{antiviral}}$ is the pattern occurrence frequency within the antiviral compound set, whereas f_{ChEMBL} is the default pattern occurrence frequency within the 1.2 million ChEMBL compounds.

On the other hand, it is interesting to assess whether a pattern is privileged by compounds associated with a given antiviral class, with respect to its occurrence frequency among all antivirals. This class specificity, *Csp*, can be written as

$$Csp(RP@C) = \frac{f_{\text{positives of class C}}(RP)}{f_{\text{antiviral}}(RP)} \quad (4)$$

with $f_{\text{positives of class C}}$ being the *RP* occurrence frequency within the subset of positives of the class *C*. A number of patterns and scaffolds were found to be prevalent with both the antiviral status in general and specific classes in particular. The most prominent privileged responsibility patterns (PRP) were selected for an in-depth discussion if it was seen to occur at least 20 times within the positives of either activity class, and both *Asp* and *Csp* (for at least one of the seven *C*) reached values of 10 or more.

2.5. GTM-Based Antiviral Propensity Model Validation. A GTM colored by node-specific predominance of positives vs negatives of a given antiviral class may be used as a predictor of the category to which a novel, so far unreported compound is most likely to belong. This is achieved by positioning the novel compound on the colored map, and reading out the locally predominant class at its residence point. The three antiviral maps built in this work were challenged to assign compounds of known antiviral class association, but not accounted for at the training stage. This external model validation involved 30 compounds that are active against HIV,³⁷ HCV,^{38,39} HSV⁴⁰ and influenza A.^{41–44} Test set compounds are considered to be predicted positives of a particular class if at least two of the three maps position them in positive-dominated class landscape zones.

3. RESULTS AND DISCUSSION

3.1. Optimal Antiviral Maps. The three top performing antiviral maps—in terms of simultaneously discriminating positives from negatives, for all the seven antiviral classes, in a 3-fold cross-validated prediction run—are depicted in Table 2. They were chosen to represent three different chemical spaces, have various sizes, but are all successful in solving the above-mentioned cross-validated discrimination problem, meaning that there is no unique recipe to capture the chemical information associated with antiviral activities.

All the maps manage to highlight specific chemical space zones associated with each of the considered classes, with no balanced accuracy scoring below a respectable 0.77. Distinction between the positives and negatives is most difficult for the lentivirus class, which is also the richest one in terms of associated positives. One may find in the archive file *XVstat.tar* provided in the Supporting Information the detailed training and cross-validation statistics for each of the thrice repeated 3-fold cross-validation attempts, including the fraction of correctly classified positives (“sensitivity”) and negatives (“specificity”) that compose the BA score.

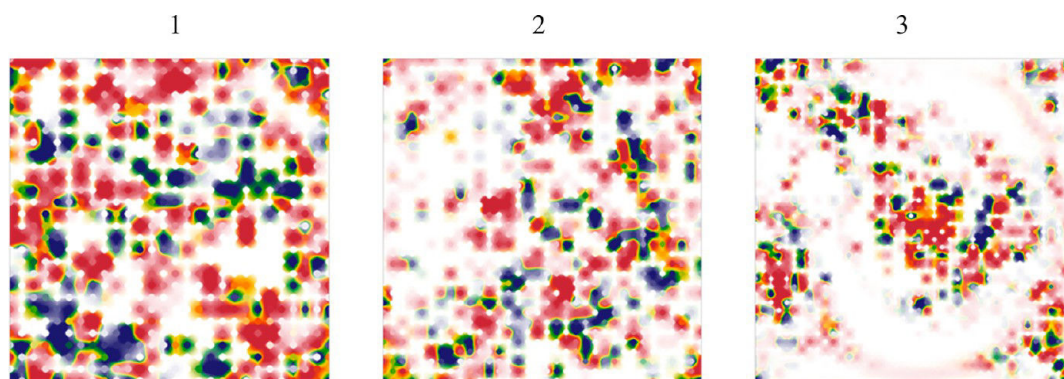


Figure 3. Classification landscape of Lentivirus-positives (class “2”, blue) vs Lentivirus-negatives (class “1”, red) on the three antiviral maps. Maps are numbered according to Table 2.

Validation of the GTM building process is conceptually more complex than the one of a typical regression or classification model, because it includes several distinct steps. First, manifold construction is totally unsupervised, already published results⁴⁵ show that being part of the frame set serving for manifold fitting is not enhancing the quality of prediction of such compounds. Next, a given manifold needs to be “colored” by a property, using a training set of compounds, and the resulting property or class landscape may serve for prediction of external compounds. By default, training and external compounds may be obtained by splitting the complete pool of available structure–property information into (typically) 2/3 for training and 1/3 for external prediction. Iteratively, each tier plays the role of test set, being subject of antiviral (positive/negative) status prediction, by projection on the map colored by the other two tiers. This test set is external, because it never contributed to color the underlying map. The data set is shuffled and the procedure is repeated three times. Reshuffling and repeating ensures that the prediction outcomes are not biased by any peculiarly favorable regrouping of compounds in test and training tiers. This “aggressive” triplicated 3-fold cross-validation (XV) adopted in this work is simply the more rigorous alternative to classical external testing on a single test set. In terms of computational effort, triplicated 3-fold XV amounts thus to nine GTM “coloring”/prediction cycles, so takes roughly the same time as a 9-fold classical XV, all while being both a much more challenging exercise—because it minimizes the information effectively used for model learning and thus maximizes the opportunities for misprediction.

Triplicated 3-fold XV is the source of map goodness (“fitness”, in the evolutionary context). Therefore, the entire available antiviral SAR information extracted from ChEMBL was used for map selection. The selected maps above are the maps that maximize predictive power, in terms of separation propensities of the considered antiviral classes, in the context of aggressive, triplicate 3-fold cross-validation. It is thus justified to ask the question whether there is a risk of “overfitting” by throwing the entire SAR information into the map selection process. In this context, “overfitting” means that the maps perform well on the current SAR data only because they were selected to perform well with respect to them. Allegedly, they may not perform as well on different antiviral compound collections. However, we do not dispose of an independent SAR data set of comparable size and richness to challenge directly this issue. Nonetheless, we do have the answer to the alternative, but equivalent question: could maps that were not selected on the basis of the present SAR data succeed with its

classification, at propensity levels comparable with the ones reported in Table 2? The answer, constituting a highlight of previous work,⁴⁵ dedicated to search of “universal” GTM models of maximal generality, is clearly “yes”. The current antiviral SAR sets were used to challenge maps that were built and selected on the basis of completely unrelated frame and color/selection sets, designed to represent the entire ChEMBL drug-like space. In that work, there was no focus on any particular property-related chemical space zone. Or, this validation was successful and returned balanced accuracies between 0.84 and 0.69, which is still significantly above the randomness level. It is thus clear that GTM models build on frame sets properly encompassing the relevant chemical space and based on relevant descriptors will be able to support antiviral class separation (note: frame sets in the cited “universal” maps did not include antiviral compounds, but were representative of the ChEMBL chemical space). In other words, GTM models built on the basis of a relevant sample of compounds (frame set) and selected because of similarity principle-compliance with respect to a series of properties typically remain similarity-principle compliant when challenged to map not yet seen compound associated with novel properties, even properties of completely different nature (systemic antiviral action, by contrast to enzyme/receptor inhibition).

Therefore, the choice to exploit all the available SAR data for map selection has been a deliberate one, chosen by contrast to previous work, where model “universality” and ability to generalize to external properties was the paramount focus. As expected, balanced accuracies reported in Table 2 exceed the ones achieved by the “universal” maps for which selection was not guided by antiviral SAR data. Although the “universal” underfitted model might be the one with better extrapolation propensity (larger applicability domain), the present work is centered on the audit of available antiviral SAR, by visualization and privileged pattern detection, which calls for an intensive exploitation of all the SAR data at hand. This notwithstanding, current maps successfully participated in prospective virtual screening followed by experimental validation (publication in preparation).

Figure 3 represents the classification landscape of the lentivirus-positives by contrast to the rest of the antiviral compounds, and it clearly displays the multiple blue zones in which lentivirus-positives “cluster” together on the map. The existence of such zones is a consequence of the high balanced accuracy values (a map with no discriminating power would be entirely colored in yellow-green, and return a balanced accuracy

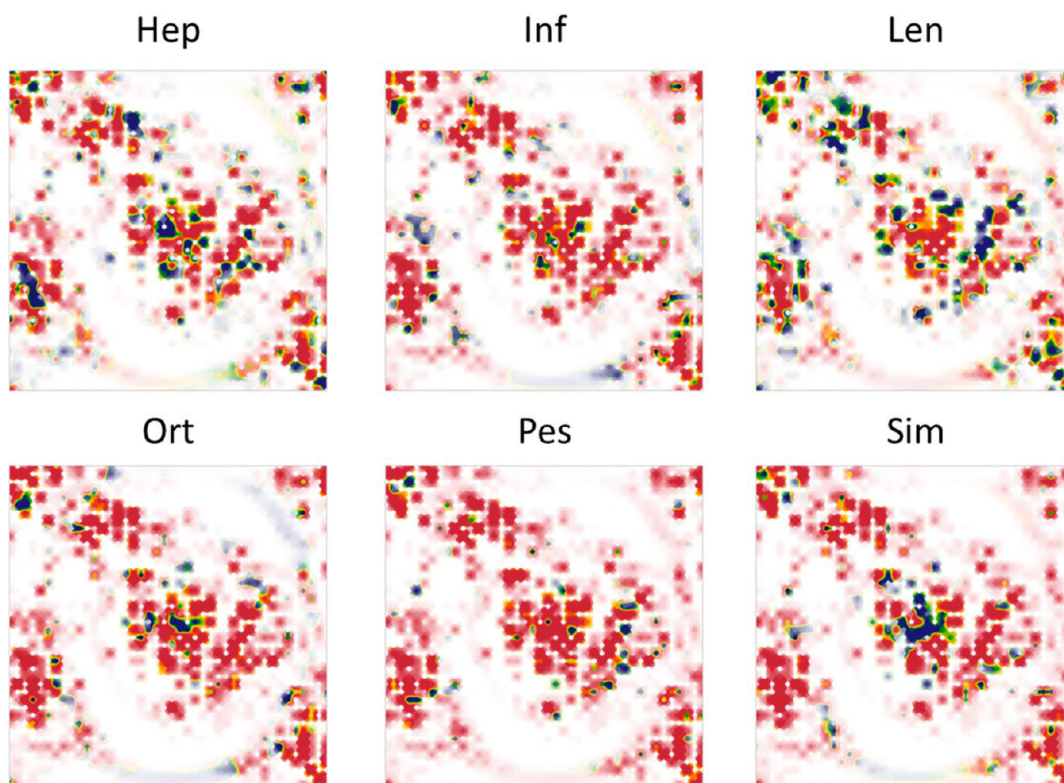


Figure 4. Classification landscapes for six of the seven virus classes, on map 3 (enterovirus, left out, is the smallest group mapping on few distinct nodes).

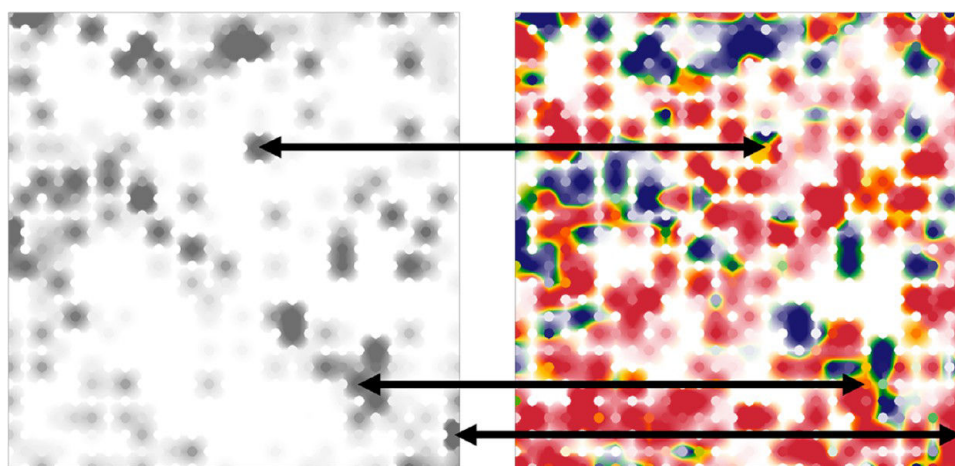


Figure 5. Density trace of Hep-positives (left) versus classification landscape of Hep-positives, as rendered by map 1. Arrows highlight areas that are densely populated by Hep-positives, but are dominated by antivirals of different classes.

score about 0.5). However, these multiple zones are scattered all over the relevant chemical space, signaling that lentivirus-positives for a large and very diverse collection of different compounds, targeting different antiviral mechanisms. Unfortunately, the ChEMBL database does not report a sufficiently large series of viral enzyme inhibition tests, which might have helped to assigning the various “lentivirus islands” on the map to different mechanisms of action (if viral target inhibitors would be found to reside within above observed islands). Failure to retrieve sufficient *in vitro* activity data against virus targets (including well-known proteins such as HIV protease and reverse transcriptase) from ChEMBL shows that the present-day antiviral compound research has been driven

forward mostly by phenotypic tests. Specific assays aimed at understanding the interactions with target proteins were realized only for few validated leads or short MedChem series of analogues—but such compound sets are small, biased and do not support global statements with respect to the entire antiviral chemical space.

The other interesting observation from Figure 3 is that the increase of map resolution (grid size) translates to an increase of the number of marginally populated nodes, and not to a better distribution of the antivirals over more nodes. This is not surprising, because the increase of the map size was not followed by an increase in discriminating power, suggesting that

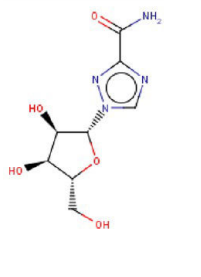
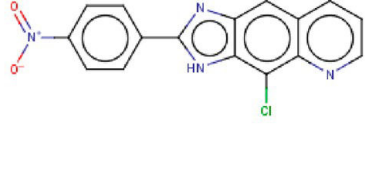
the smaller map 1 is already sufficiently large to accommodate the chemical diversity spanned by the antiviral compounds.

Figure 4 is a comparative display of the classification landscapes for six out of seven antiviral classes of compounds on the map 3 (Table 2). All these represent a same global compound set—the entire antiviral set—in which the space zones dominated by each of the six classes are, alternatively, highlighted. Positives of every antiviral class form chemically diverse collections, but each has a rather distinct scattering pattern on the map. Compounds associated with different virus classes show clear and distinct pictorial “signatures” on the map.

When classification landscapes like in Figure 4 are matched against one-class plots (positives of a given virus class) shown in Figure 5, it is possible to evidence the chemical space areas where the positives are outnumbered in terms of normalized density. The left-hand plot in Figure 5 represents the density trace of the Hep positives. If these would exclusively occupy chemical space zones void of, or sparsely populated by any other antivirals, then all the high-density areas on the left should match blue, Hep-dominated areas in the classification landscape right. This is mostly true; otherwise, no high balanced accuracy score could have been reached, but not always. Visually, it is easy to pinpoint the areas in which significant subsets of Hep-positives are outnumbered by other antiviral compounds (three such spots were highlighted).

There may be two alternative explanations for the existence of such mismatches: the pessimistic one is that the limit of accuracy of the GTM model is attained, whereas the optimistic one would be the claim that therein found Hep-negatives are actually not yet discovered actives. The latter is indirectly supported by the actual existence of promiscuous compounds, known to belong to both the Hep and other classes, as in Table 3. They are not the ones contributing to the dilution of Hep-

Table 3. Examples of Promiscuous Antivirals

	
CHEMBL1643(Ribavirin) Activity: Hep, Inf, Pes, Sim	CHEMBL1940452 Activity: Hep, Pes, Sim

positive population in the right-hand plot of Figure 5 (when in several classes, compounds are counted as “blue” in “class versus remainder of antiviral” plots), but they are indirect evidence in favor of the possibility of promiscuous molecules.

More details can be provided by constructing specific “class versus class” landscapes, by contrast to the above “one class versus remaining antivirals”. Plots of Hep-positives (set as class “2”, blue) versus the positives of the six other classes (each in the role of class “1”, red) may reveal which are the other classes that overlap with the problematic areas in Figure 6. If a class does not interfere, the areas should be Hep-dominated (blue): Pes-positives are absent from the both encircled areas, whereas Ent-positives are absent from the “south-east” area only. The corner zone, not encircled, seems to have little specificity, and

harbors structures of all the seven classes—probably a “garbage” area receiving compounds that are not closely approached by the manifold. Eventually, even though ChEMBL compounds labeled as nonantivirals (status “0”) were never used in the map selection process, Figure 7 below clearly illustrates that the maps are nevertheless well able to distinguish these from the antiviral molecules. This is not a trivial result, for unlike the “universal” maps reported in a previous work,⁴⁵ the frame sets used for antiviral map building failed to include major drug-like categories such as GPCR binders. On low-resolution maps like map 1, these various “novel” chemotypes not covered by the frame set seem to collapse into a few very high-density nodes (7 nodes have cumulated responsibilities above 25 000 compounds each). By contrast, in higher-resolution map 3, the nonantiviral compounds seem to map onto the zones that were left largely empty by the antiviral molecules.

3.2. External Test Set Validation. The best external validation of predictive models will always remain the prospective virtual screening of compound databases, followed by experimental testing and discovery of novel actives. The herein developed GTM models were instrumental in the discovery process of novel antiviral chemical entities, in a collaborative study involving medicinal chemists and virologists, and which will be published elsewhere. The second best way to validate models is *a posteriori* application to external known actives from sources other than the ones used for training, thus mimicking at best the virtual screening context.

Unfortunately, the quest for genuinely “external” validation data in the above sense was of rather limited success. The additional data used to challenge the maps in an external antiviral class prediction exercise consisted of 10 anti-Influenza A virus, 2 antientivirus, 5 antihepacivirus and 2 antisimplexvirus compounds. Except for the influenza A subset, these numbers are too scarce for robust validation (a state of fact showing how difficult it may be, in practice, to find experimental data not yet part of ChEMBL), but prediction was attempted nevertheless. Results were excellent for the influenza A subset, where 9 out of 10 candidates were correctly recognized as positives. Also, both antientivirus compounds have been recognized, however none of the antisimplex or antihepacivirus candidates were predicted positive.

3.3. Privileged Pattern Analysis. Note that the empirical offset of 0.9 in eq 2 was chosen such as to set the empirical border between “irrelevant” and “marginally relevant” nodes at 0.01. This empirical choice is in agreement with our general experience with the GTM tool, which shows that nodes with responsibilities above 1% are relatively rare events, deserving to be distinguished from the typical “unpopulated” nodes featuring much lower tail values of the responsibility distributions. However, in the following, this peculiar choice is of no relevance. The most relevant privileged responsibility patterns (PRPs), satisfying the empirical criteria given in section 2.4, refer to single map nodes, monopolizing 100% of the responsibility distribution of associated compounds. In this specific case, PRPs may be thought of as “privileged map nodes”. These nodes were highlighted in Figure 7, in the context of displaying the generic antiviral compound space by contrast to the rest of the ChEMBL. The key PRPs are both specific to antivirals versus nonantiviral compounds, and, furthermore, specific to some antiviral classes as by contrast to the remainder of antiviral chemical space. As such, they

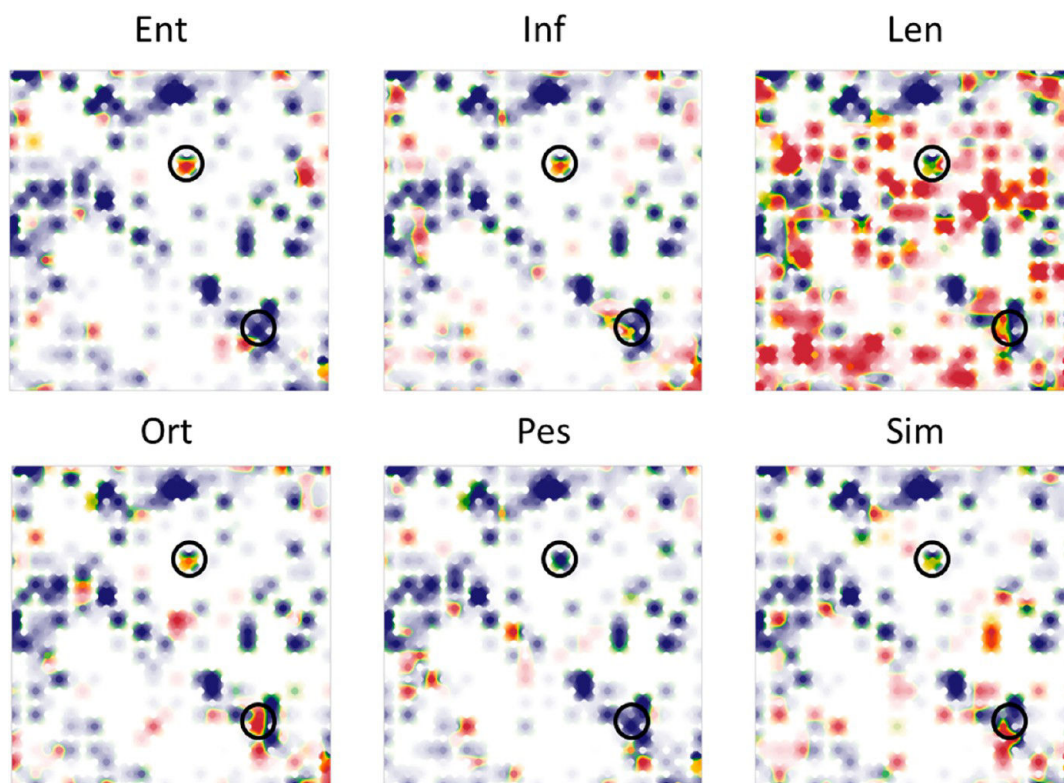


Figure 6. Pairwise classification landscapes built for Map 1 confronting hepacivirus positives (dominant in blue areas) with positives of the six other antiviral classes, respectively. Encircled zones correspond to two of the problem spots highlighted in Figure 5, with the third—the southeast corner—seemingly attracting positives of all the classes.

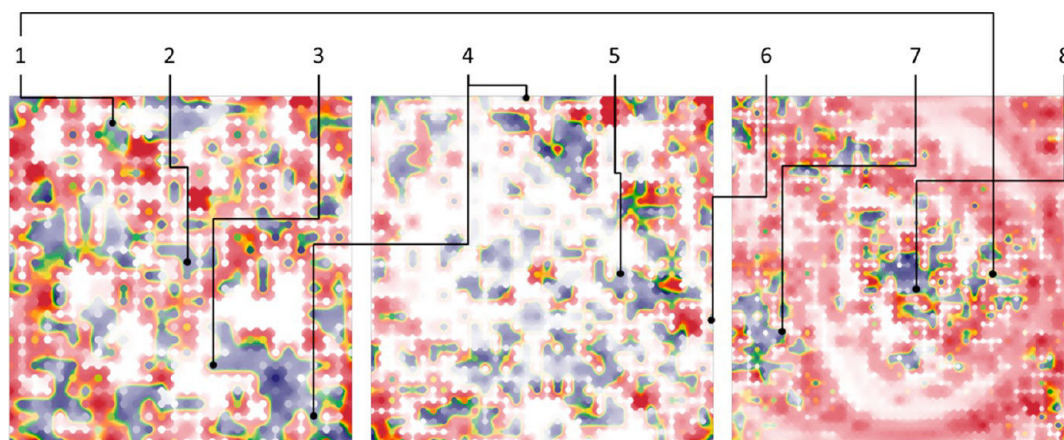


Figure 7. Classification landscapes of antiviral compounds, all classes confounded and labeled “2” (blue) versus nonantiviral ChEMBL molecules (status “0” in the original classification, here acting as class “1”, in red). Highlighted nodes on the maps (#1,2,3 from left to right, as in Table 2) correspond to the single-node PRPs harboring the PSMs shown in Table 4.

consistently fall within blue, antiviral chemical space zones—but do not represent maximal density areas.

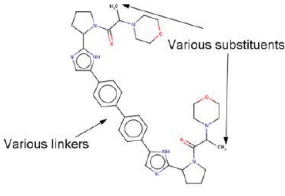
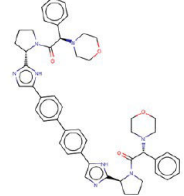
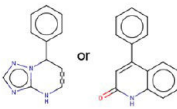
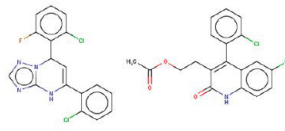

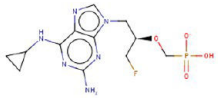
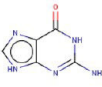
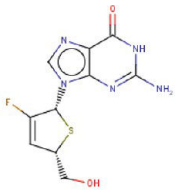
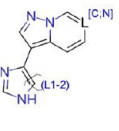
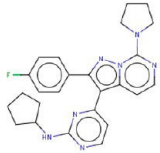
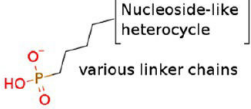
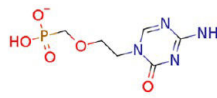
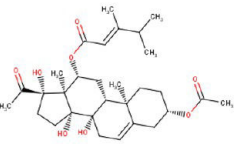
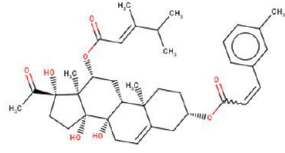
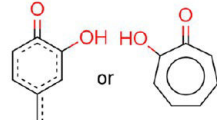
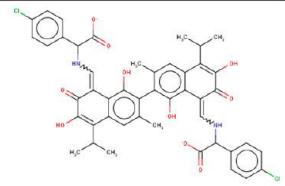
The privileged structural motifs (PSMs) of compounds associated with each PRP were determined by visual inspection. Note that a same PSM might be independently discovered as underlying structural motifs of PRPs on different maps (PSM number, same as in Table 4 are connected to nodes from different maps in Figure 7).

PSM 1 was selected because privileged by the pestivirus class, and was convergently discovered in PRPs of both maps 1 and 3. This node harbors 36 of the 412 pestivirus-positives, which brings its $Csp(Pes)$ value to 17. However, it is dominated by

120 anti-HCV compounds which are postulated to be HCV nonstructural proteins inhibitors.^{22,23} Because, however, the pool of hepacivirus-positives is intrinsically larger (>5300), the 120 compounds only account for a hepacivirus-class Csp of 4.5. Note that 34 compounds of the 36 *Pes*-positives are actually labeled as both *Hep* and *Pes*. Most compounds (110 out of 120), such as Daclatasvir, are imidazolylpyrrolidines.

As this example clearly shows, privileged status of a RP with respect to a class does not mean that the given class is necessarily the best represented within that RP—it means that a relative majority of compounds from that class match the given RP. Compounds of other classes may dominate that

Table 4. Main Privileged Antiviral Structural Motifs Resulted from the Analysis of Responsibility-Based Patterns^a

#	Class	Privilege Structural Motif	Representative molecules
1	<i>Pes</i>		
2	<i>Ort</i>		
3	<i>Sim</i>		
4	<i>Sim</i>		
5	<i>Sim</i>		
6	<i>Sim</i>		
7	<i>Ort</i>		
8	<i>Inf</i>		

^aLocation of the structures on the maps is shown in Figure 7.

RP—yet, if they represent less significant fractions of their respective classes, this RP may be less privileged with respect to the latter. The example also suggests that the arbitrary factor of 10 chosen to pick the most “extreme” cases of privileged patterns for discussion is far too restrictive: useful insight may be gained from patterns at lower values. Note that for *Hep* and *Len*, *Csp* scores of 10 are impossible due to their high occurrence in antiviral data set. Even if a given RP would exclusively occur within one of these classes, $f_{\text{positive}} = (X \text{ occurrences}/F \text{ class members})$ reported to $f_{\text{antiviral}} = (\text{same } X \text{ occurrences}/A \text{ antivirals})$ cannot exceed A/F , a ratio well below

10 for either of *Hep* and *Len*. The above-mentioned *Csp*(*Hep*) at 4.5 is virtually equal to the absolute maximum $A/F = 24629/5320 = 4.63$ achievable within this data collection. The absence in Table 4 of PRPs specifically dedicated to the two main antiviral classes is not a problem—they were detected during the analysis, but were not picked for the present proof-of-concept discussion of PRPs as direct sources of privileged structural motifs.

PSM 2, privileged by the Orthohepadnavirus class with a *Csp* exceeding 30, is harboring a structurally diverse set of compounds possessing different activities, such as antimeasles³²

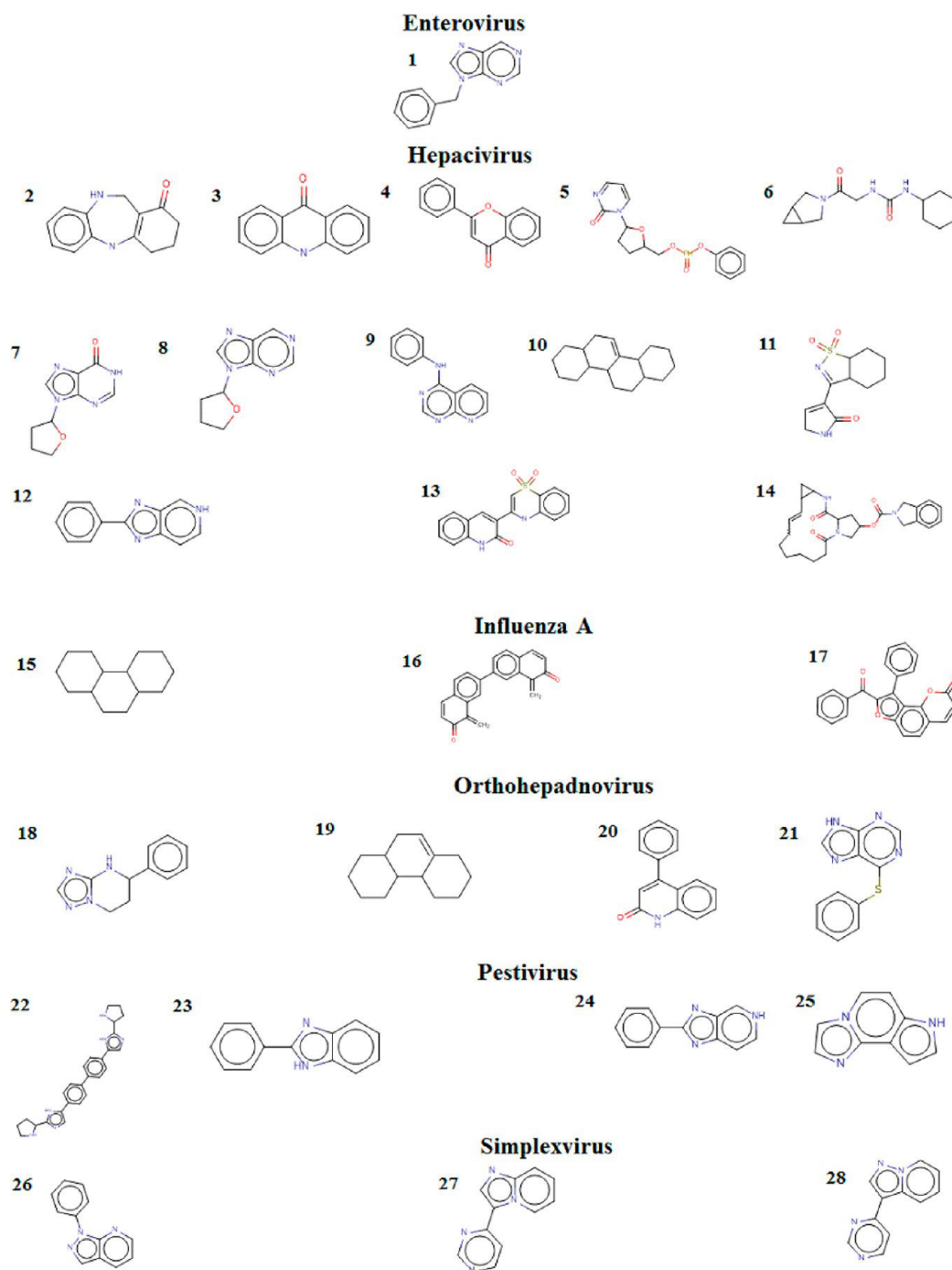


Figure 8. Scaffolds for antiviral compounds, extracted for each antiviral class by Scaffold Hunter.

and anti-HBV(HBsAg inhibitors).³³ The majority of them are variations of either of the two distinct but similar scaffolds shown in Table 4. Hence, this PSM regroups a pair of similar scaffolds that are allegedly interchangeable options in antiviral compound design, as the knowledge extracted by generative topographic mapping seems to suggest.

Both PSMs 3 and 4 (privileged by simplexvirus) consist of nucleoside-based analogs with a broad spectrum of antiviral activity. This is a case where the highlighted “structural motif” coincides with the classical definition of privileged scaffolds. The first PSM is mainly seen in anti-HSV²⁵ and anti-HIV²⁴ compounds, whereas the second, mainly anti-HSV-oriented,^{26,27} includes popular drugs Ganciclovir and Aciclovir.

PSM 5 (privileged by simplexvirus) comprises antivirals with various polyheterocyclic systems.²⁸ This pattern regroups a series of very close but distinct scaffolds, differing in terms of ring size (five- versus six-membered) and the positions of aromatic N atoms on the otherwise conserved scaffold graph. This example shows that responsibility patterns are able to spontaneously regroup closely related scaffolds.

PSM 6 regroups various antisimplexvirus nucleotide mimics, with various heterocycles (including, but not restricted to, the natural purines and pyrimidines) linked via a linear chain (mainly hydrophobic, occasionally including an ether group) to a phosphate group. Clearly, on one hand PSM 6 cannot be reduced to any single scaffold, whereas, on the other, it is not

solely defined by the scaffold. Representatives of this class also display a broad spectrum of activity, particularly against HIV⁶¹ and herpesviruses.⁶²

PSM 7 (Orthohepadnavirus) translated into a very homogeneous family of steroid compounds, such as caudatin and its derivatives, which originally come from natural sources and were found effective against HBV.³⁴ It is noteworthy that in this case the actual definition of the privileged motif is very precise: more specific than the mere scaffold structure. Not only the scaffold *per se* but also some of its “ornaments” appear to be conserved throughout the group. Note that the retrieved pattern is chiral, albeit chirality is ignored by the used molecular descriptors. This should not be interpreted as some prediction of the required chirality, but simply as an observation that the current motif systematically appears under this single stereochemistry in the database, which is not surprising within a series of chemically modified natural products. Would ChEMBL have contained different stereoisomers of this moiety, those would have been mapped onto the same node, and, if not listed among known antivirals, would have “eroded” the privileged status of the pattern. This clearly shows (a) the intrinsic limitation of any 2D descriptor-based analysis and (b) that a privileged status is often not a reflection of the intrinsic preference of the target for that moiety, but a mere bias due to absence of “negative” counterexamples featuring that pattern.

PSM 8 emerges as privileged of both influenza A and Orthohepadnavirus classes and covers a rather diverse structural family, most of which (but not all) have in common the benzoquinone dimer highlighted in Table 4, embedded in a large variety of chemical contexts. Counterexamples not featuring this dimer core contain a single quinone moiety or, alternatively, a tropolone core. Many of the species appear as negatively charged at physiological pH, either due to ionization of rather ubiquitous phenol groups, or due to the presence of sulfonate and carboxylate anions. Albeit this motif does not seem allow any simple definition in terms of common scaffolds, regrouping these—putatively redox-active—quinone/polyphe-nols together does make perfect chemical sense. It is an example of a fuzzy but meaningful motif that could not have been highlighted as such by substructure mapping. The compounds display antiviral activity against HIV,^{29,31} HBV³⁰ and influenza A.³¹

Thus, the analysis of PRP-based compound clusters turned out to be a tool of high versatility, because it does not rely on any preconception on the nature of the structural motif to look for. Sometimes, the PSM found to characterize the given subset of antivirals actually happens to coincide with the presence of a privileged antiviral scaffold—nucleosides, notably. However, in some cases the actual motif may be more finely tuned than simple scaffold presence—the privileged structure may be a specifically substituted scaffold, not any occurrence thereof. By contrast, sometimes the common characteristic of an antiviral compound subset may be too fuzzy to pinpoint in terms of specific substructures, all while making nevertheless perfect chemical sense, as was the case of the rather diverse (redox-active?) phenol/quinone species, or the series of rather diverse nucleotide mimics. Note that some PSM are being specifically “discovered” by several of the maps, each independently allotting a node for harboring broadly the same subset of structurally related compounds. By contrast, others are specifically highlighted by only one of the three maps, which are thus able to provide complementary perspective overviews of the antiviral space.

It is interesting to note that virtually all highlighted patterns have strong relatedness to classes of natural compounds: peptides, nucleosides, sterols and polyphenols. Natural compounds or derivatives thereof appear to be privileged in antiviral research, perhaps more than in other “more rational” branches of drug design. This trend is spontaneously highlighted by the GTM-driven analysis.

In Supplementary M, one will find a list of many more interesting RPs, selected at less strict criteria ($Asp > 1$; $Csp > 1$; at least 10 positives within the antiviral class) in the three files *privPat*{1,2,3} corresponding to the three maps. RPs are rendered by formatted strings like “/NODE1:RP(NODE1)/NODE2:RP(NODE2)/...” providing a slash-separated list of relevant nodes with RP values above zero, associated by “:” with the actual RP values. The compounds associated with a given RP pattern can be found by searching this pattern in the second column of the provided *AVmap*{1,2,3}*Resp.rpat* files (again, one for each map), where column one is simply the positional ID of compounds, i.e., the line number at which they can be found in the structure-ID file *StdChEMBL.smi_chid*.

3.4. Scaffold Analysis. The main problem with scaffold analysis is the ambiguity of the scaffold concept: it can be a common substructure including only intracyclic bonds, or also allowing rings to be interconnected by an acyclic linker of arbitrary length, it can be taken as the bare polycyclic graph, ignoring the nature of heteroatoms, *etc.*⁶⁰ In this work, only scaffolds with at least 3 cyclic fragments—fused or interconnected by acyclic linkers—were selected for further consideration, if they were present within at least 20 positive structures of an antiviral class. Some scaffolds may represent substructures of larger ones, and were not discarded.

When the privileged status of “naked” scaffolds was assessed, it was seen that among the 28 checked scaffolds (see Figure 8), only 5 correspond to the criteria of being “privileged”: these are numbers 16, 17, 22, 25 and 28. Some of these were already discussed, because they are present within the responsibility-driven compound clusters in Table 4. Even scaffolds that codefine fuzzier PSMs in combination with different, related substructures may nevertheless score high *Asp* and *Csp* values—classical analysis would have highlighted them as privileged, whereas in fact they are only peculiar “incarnations” of a broader motif as highlighted previously. Such examples include scaffold #16, a frequent representative of PSM 8, coexisting next other various cores of hydroxylated quinone or tropolone type. Similarly, scaffold #22 is one peculiar substructure appearing in PSM 1, and the same applies for scaffold #28 with respect to PSM 5. Scaffold #25 compounds are active against BVDV-1,³⁵ whereas scaffold #17 is privilegedly encountered in positives of the influenza A class.

4. CONCLUSIONS

This work represents an audit of antiviral structure–activity data in the public database ChEMBL, using cheminformatics tools and in particularly aimed at showing how the rather recent technique of generative topographic mapping (GTM) may be used to render rationally, visualize intuitively, model and predict antiviral activities as a function of compound structure. More precisely, targeted goals were to

- Curate and standardize structure-antiviral activity in ChEMBL;
- Provide an association of individual structures with seven broad virus classes, transcending the numerous and

- diverse antiviral test protocols, thus building large and robust structure-class training sets that are perfectly suited for categorical model building;
- Build dedicated GTMs that optimally discriminate between the above-mentioned classes, and to use these for visualization of the antiviral chemical space in the context of the entire ChEMBL compound collection, and of the specific space zones allotted to the specified antiviral classes;
 - Use generated maps to focus attention on specific responsibility patterns—corresponding to specific locations on the map—that appear as “privileged” by certain antiviral classes;
 - Understand how these responsibility patterns relate to the structural features of molecules seen to cluster together, and compare insights that can be gained from privileged responsibility patterns to the classical scaffold-based “privileged structure” analysis.

A first important conclusion is that so-far available public data is largely insufficient in order to allow a rigorous buildup of an actual structure–antiviral activity profile. Experimental information is sparse, as no compound has been systematically tested against all the virus strains. Therefore, data fusion of individual assay results, in order to assign generic antiviral class-based membership labels is, so far, the only way we found to extract statistically exploitable training sets supporting the attempted analysis of antiviral chemical space.

Despite the intrinsic uncertainty of used class labels (a “negative” may be wrongly assigned, because so far not tested on that virus class), GTMs successfully separated positives from negatives, with 3-fold cross-validated balanced accuracy scores of 0.8–0.9. External validation—challenging the maps to detect antiviral compounds outside of the ChEMBL database, and not used at map growing stage—was a partial success, especially with respect to Influenza A compounds (20 out of 21 were recognized as such, after mapping).

Visually, the separation into classes, each preferentially mapping to other areas on the map, can be clearly observed. For example, it is straightforward to visualize the relative positioning of the positives associated with any two antiviral classes, observing their potential overlap zones where “promiscuous” compounds may reside. This may enable antiviral compound repositioning, if the compound is seen to reside in an overlap zone involving both its so-far targeted virus class and another, yet unassessed virus class.

Detection of responsibility patterns (herein matching well-defined nodes on the maps) that are “privileged” by any antiviral class (in the sense of occurring more often than expected on a random basis within the associate “positive” compounds) was proven to represent a powerful generalization of the classical search for “privileged structures”, a central paradigm in modern medicinal chemistry. Establishing the “privileged” status of a structural motif is a simple statistical exercise—however, a medicinal chemist is facing a virtual infinity of possible motifs (substructures, connected or disjointed subgraphs, pharmacophore patterns, molecular field patterns) for which the privileged status should be assessed. They focus their attention on scaffolds. The use of GTM technology, however, provides a natural answer to the key question “what is the nature of the privileged structural features?” Therefore, responsibility pattern analysis is a

powerful application of GTM technology, able to spontaneously adjust to the correct “resolution” needed:

- At scaffold level. In some cases, privileged responsibility patterns are seen to gravitate around a common scaffold, which could have been picked as “privileged” by a classical analysis.
- Coarser than scaffold level (such as the large and diverse family of polyphenols/quinones/tropolones, where the common trait seem to be the redox-competent functional groups rather than any specific scaffold).
- Finer than scaffold level, the virus group turns out to privilege not the scaffold *per se*, but a specifically substituted scaffold, as exemplified by the substituted steroid core of Table 4. In this case, the scaffold alone might not even be recognized as privileged, and the insight would have been lost in classical analysis.

The nonlinear nature of GTM models coupled to the evolutionary optimization—including the selection of best suited molecular descriptor schemes, bound to capture relevant structural information—allows to automatically tune in to the best resolution level needed to capture privileged structural characteristics in general, rather than predefined scaffolds that may or may not match the trend present in experimental data. Some privileged structural motifs were being reproducibly highlighted by several of the maps, each independently allotting a node for harboring roughly the same subset of structurally related compounds. By contrast, other chemical features are specifically highlighted by only one of the three considered antiviral maps. These different mapping schemes, based on different molecular descriptor sets, are thus partly convergent and partly complementary in terms of the light they shed on the antiviral space.

■ ASSOCIATED CONTENT

📄 Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jcim.6b00192.

Further document with supplementary tables cited in the text, as well as large data files (Unix text) meant for computer-aided exploitation and querying; the file names and their content has been explained in the main text, in the concerned paragraphs (ZIP).

■ AUTHOR INFORMATION

Corresponding Authors

*A. Varnek. E-mail: varnek@unistra.fr.

*D. Horvath. E-mail: dhorvath@unistra.fr.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

K.K. is thankful to French embassy in Ukraine for his Ph.D. fellowship.

■ ABBREVIATIONS

GTM, generative topographic map; (P)RP, (privileged) responsibility pattern; PSM, privileged structural motif

■ REFERENCES

(1) Cleaveland, S.; Fevre, E. M.; Kaare, M.; Coleman, P. G. Estimating Human Rabies Mortality in The United Republic of

Tanzania from Dog Bite Injuries. *Bull. World Health Organ.* **2002**, *80*, 304–310.

(2) Shrestha, S. S.; Swerdlow, D. L.; Borse, R. H.; Prabhu, V. S.; Finelli, L.; Atkins, C. Y.; Owusu-Edusei, K.; Bell, B.; Mead, P. S.; Biggerstaff, M.; Brammer, L.; Davidson, H.; Jernigan, D.; Jhung, M. A.; Kamimoto, L. A.; Merlin, T. L.; Nowell, M.; Redd, S. C.; Reed, C.; Schuchat, A.; Meltzer, M. I. Estimating the Burden of 2009 Pandemic Influenza A (H1N1) in the United States (April 2009–April 2010). *Clin. Infect. Dis.* **2011**, *52* (Suppl. 1), S75–S82.

(3) De Clercq, E. Dancing with Chemical Formulae of Antivirals: A Panoramic View (Part 2). *Biochem. Pharmacol.* **2013**, *86*, 1397–1410.

(4) Domingo, E.; Holland, J. J. RNA Virus Mutations and Fitness for Survival. *Annu. Rev. Microbiol.* **1997**, *51*, 151–178.

(5) ChEMBL Database, version 19, doi: [10.6019/CHEMBL.database-19](https://doi.org/10.6019/CHEMBL.database-19) (accessed November, 2014).

(6) Poduch, E.; Bello, A. M.; Tang, S. H.; Fujihashi, M.; Pai, E. F.; Kotra, L. P. Design of inhibitors of orotidine monophosphate decarboxylase using bioisosteric replacement and determination of inhibition kinetics. *J. Med. Chem.* **2006**, *49*, 4937–4945.

(7) Bishop, C. M.; Svendsen, M.; Williams, C. K. I. Developments of the Generative Topographic Mapping. *Neurocomputing* **1998**, *21*, 203–224.

(8) Bossart-Whitaker, P.; Carson, M.; Babu, Y. S.; Smith, C. D.; Laver, W. G.; Air, G. M. Three-dimensional Structure of Influenza A N9 Neuraminidase and Its Complex with the Inhibitor 2-deoxy-2,3-dehydro-N-acetyl Neuraminic Acid. *J. Mol. Biol.* **1993**, *232*, 1069–1083.

(9) Gregoriades, A. The Membrane Protein of Influenza Virus: Extraction from Virus and Infected Cell with Acidic Chloroform-Methanol. *Virology* **1973**, *54*, 369–383.

(10) Berthold, M. R.; Cebon, N.; Dill, F.; Gabriel, T. R.; Kötter, T.; Meinel, T.; Ohl, P.; Sieb, C.; Thiel, K.; Wiswedel, B. KNIME: The Konstanz Information Miner. In *Data Analysis, Machine Learning and Applications*; Preisach, C., Burkhardt, H., Schmidt-Thieme, L., Decker, R., Eds.; Springer: Berlin, Germany, 2007; pp 319–326.

(11) Kumar, R.; Nath, M.; Tyrrell, D. L. J. Design and Synthesis of Novel 5-Substituted Acyclic Pyrimidine Nucleosides as Potent and Selective Inhibitors of Hepatitis B Virus. *J. Med. Chem.* **2002**, *45*, 2032–2040.

(12) Lin, T. S.; Zhu, J. L.; Dutschman, G. E.; Cheng, Y. C.; Prusoff, W. H. Syntheses and Biological Evaluations of 3'-Deoxy-3'-C-Branched-Chain-Substituted Nucleosides. *J. Med. Chem.* **1993**, *36*, 353–362.

(13) DeGoey, D. A.; Randolph, J. T.; Liu, D.; Pratt, J.; Hutchins, C.; Donner, P.; Krueger, A. C.; Matulenko, M.; Patel, S.; Motter, C. E.; Nelson, L.; Keddy, R.; Tufano, M.; Caspi, D. D.; Krishnan, P.; Mistry, N.; Koev, G.; Reisch, T. J.; Mondal, R.; Pilot-Matias, T.; Gao, Y.; Beno, D. W.; Maring, C. J.; Molla, A.; Dumas, E.; Campbell, A.; Williams, L.; Collins, C.; Wagner, R.; Kati, W. M. Discovery of ABT-267, a Pan-Genotypic Inhibitor of HCV NS5A. *J. Med. Chem.* **2014**, *57*, 2047–2057.

(14) Tonelli, M.; Boido, V.; Canu, C.; Sparatore, A.; Sparatore, F.; Paneni, M. S.; Fermeglia, M.; Prici, S.; La Colla, P.; Casula, L.; Ibba, C.; Collu, D.; Loddo, R. Antimicrobial and cytotoxic arylazoenamines. Part III: antiviral activity of selected classes of arylazoenamines. *Bioorg. Med. Chem.* **2008**, *16*, 8447–8465.

(15) Giliberti, G.; Ibba, C.; Marongiu, E.; Loddo, R.; Tonelli, M.; Boido, V.; Laurini, E.; Posocco, P.; Fermeglia, M.; Prici, S. Synergistic Experimental/Computational Studies on Arylazoamine Derivatives that Target the Bovine Viral Diarrhea Virus RNA-Dependent RNA Polymerase. *Bioorg. Med. Chem.* **2010**, *18*, 6055–6068.

(16) *Virus Taxonomy: Ninth Report of the International Committee on Taxonomy of Viruses*, ed. 1 King, A., Adams, M., Carstens, E., Lefkowitz, E., Eds.; Elsevier: 2012; p 1327.

(17) Scaffold Hunter, version 2.4.1, <https://sourceforge.net/projects/scaffoldhunter/> (accessed February, 2015).

(18) Schuffenhauer, A.; Ertl, P.; Roggo, S.; Wetzel, S.; Koch, M. A.; Waldmann, H. The Scaffold Tree - Visualization of the Scaffold

Universe by Hierarchical Scaffold Classification. *J. Chem. Inf. Model.* **2007**, *47*, 47–58.

(19) Antiviral IntelliStrat Database, version 2016, <https://www.antiviralintelistrat.com/> (accessed 2016).

(20) Steindl, T. M.; Crump, C. E.; Hayden, F. G.; Langer, T. Pharmacophore Modeling, Docking, and Principal Component Analysis Based Clustering: Combined Computer-Assisted Approaches to Identify New Inhibitors of the Human Rhinovirus Coat Protein. *J. Med. Chem.* **2005**, *48*, 6250–6260.

(21) Zhou, Z.; Khaliq, M.; Suk, J. E.; Patkar, C.; Li, L.; Kuhn, R. J.; Post, C. B. Antiviral Compounds Discovered by Virtual Screening of Small-Molecule Libraries Against Dengue Virus E Protein. *ACS Chem. Biol.* **2008**, *3*, 765–775.

(22) Shi, J.; Zhou, L.; Amblard, F.; Bobeck, D. R.; Zhang, H.; Liu, P.; Bondada, L.; McBrayer, T. R.; Tharnish, P. M.; Whitaker, T.; Coats, S. J.; Schinazi, R. F. Synthesis and Biological Evaluation of New Potent and Selective HCV NS5A Inhibitors. *Bioorg. Med. Chem. Lett.* **2012**, *22*, 3488–3491.

(23) Abdel-Magid, A. F. Halting HCV Replication with NS5A Inhibitors and NS5B Polymerase Inhibitors: Effective New Treatments of HCV Infection. *ACS Med. Chem. Lett.* **2014**, *5*, 234–237.

(24) Baszczynski, O.; Jansa, P.; Dracinsky, M.; Klepetarova, B.; Holy, A.; Votruba, I.; de Clercq, E.; Balzarini, J.; Janeba, Z. Synthesis and Antiviral Activity of N9-[3-fluoro-2-(phosphonomethoxy)propyl] Analogues Derived from N6-substituted Adenines and 2,6-Diaminopurines. *Bioorg. Med. Chem.* **2011**, *19*, 2114–2124.

(25) Prichard, M. N.; Hartline, C. B.; Harden, E. A.; Daily, S. L.; Beadle, J. R.; Valiaeva, N.; Kern, E. R.; Hostetler, K. Y. Inhibition of Herpesvirus Replication by Hexadecyloxypropyl Esters of Purine- and Pyrimidine-Based Phosphonomethoxyethyl Nucleoside Phosphonates. *Antimicrob. Agents Chemother.* **2008**, *52*, 4326–4330.

(26) Zhou, S. M.; Drach, J. C.; Prichard, M. N.; Zemlicka, J. (Z)- and (E)-2-(1,2-Dihydroxyethyl)methylenecyclopropane Analogues of 2'-Deoxyadenosine and 2'-Deoxyguanosine. Synthesis of All Stereoisomers, Absolute Configuration, and Antiviral Activity. *J. Med. Chem.* **2009**, *52*, 3397–3407.

(27) Diez-Torrubia, A.; Cabrera, S.; de Castro, S.; Garcia-Aparicio, C.; Mulder, G.; De Meester, I.; Camarasa, M. J.; Balzarini, J.; Velazquez, S. Novel Water-Soluble Prodrugs of Acyclovir Cleavable by the Dipeptidyl-Peptidase IV (DPP IV/CD26) Enzyme. *Eur. J. Med. Chem.* **2013**, *70*, 456–468.

(28) Gudmundsson, K. S.; Johns, B. A. Imidazo[1,2-a]pyridines with Potent Activity Against Herpesviruses. *Bioorg. Med. Chem. Lett.* **2007**, *17*, 2735–2739.

(29) Wang, Z. Q.; Bennett, E. M.; Wilson, D. J.; Salomon, C.; Vince, R. Rationally Designed Dual Inhibitors of HIV Reverse Transcriptase and Integrase. *J. Med. Chem.* **2007**, *50*, 3416–3419.

(30) Fan, G. T.; Li, Z. L.; Shen, S.; Zeng, Y.; Yang, Y. S.; Xu, M. J.; Bruhn, T.; Bruhn, H.; Morschhauser, J.; Yingmann, G.; Lin, W. H. O-sulfated Pyrrole Alkaloids with Anti-HIV-1 Activity, from the Chinese Marine Sponge *Iotrochota Baculifera*. *Bioorg. Med. Chem.* **2010**, *18*, 5466–5474.

(31) Yang, J.; Zhang, F.; Li, J. R.; Chen, G.; Wu, S. W.; Ouyang, W. J.; Pan, W.; Yu, R.; Yang, J. X.; Tien, P. Synthesis and Antiviral Activities of Novel Gossypol Derivatives. *Bioorg. Med. Chem. Lett.* **2012**, *22*, 1415–1420.

(32) White, L. K.; Yoon, J. J.; Lee, J. K.; Sun, A. M.; Du, Y. H.; Fu, H.; Snyder, J. P.; Plemper, R. K. Nonnucleoside Inhibitor of Measles Virus RNA-Dependent RNA Polymerase Complex Activity. *Antimicrob. Agents Chemother.* **2007**, *51*, 2293–2303.

(33) Guo, R. H.; Zhang, Q. A.; Ma, Y. B.; Huang, X. Y.; Luo, J.; Wang, L. J.; Geng, C. A.; Zhang, X. M.; Zhou, J.; Jiang, Z. Y.; Chen, J. J. Synthesis and Biological Assay of 4-aryl-6-chloro-quinoline Derivatives as Novel Non-Nucleoside Anti-HBV Agents. *Bioorg. Med. Chem.* **2011**, *19*, 1400–1408.

(34) Wang, L. J.; Geng, C. A.; Ma, Y. B.; Huang, X. Y.; Luo, J.; Chen, H.; Guo, R. H.; Zhang, X. M.; Chen, J. J. Synthesis, Structure-Activity Relationships and Biological Evaluation of Caudatin Derivatives as

Novel Anti-Hepatitis B Virus Agents. *Bioorg. Med. Chem.* **2012**, *20*, 2877–2888.

(35) Chezal, J. M.; Paeshuysse, J.; Gaumet, V.; Canitrot, D.; Maisonial, A.; Lartigue, C.; Gueiffier, A.; Moreau, E.; Teulade, J. C.; Chavignon, O.; Neyts, J. Synthesis and Antiviral Activity of an Imidazo[1,2-a]pyrrolo[2,3-c]pyridine Series Against The Bovine Viral Diarrhea Virus. *Eur. J. Med. Chem.* **2010**, *45*, 2044–2047.

(36) Whitley, R. J.; Roizman, B. Herpes Simplex Virus Infections. *Lancet* **2001**, *357*, 1513–1518.

(37) Buckheit, R. W.; Buckheit, K. W.; Sturdevant, C. B.; Buckheit, R. W. Selection and Characterization of Viruses Resistant to The Dual Acting Pyrimidinedione Entry and Non-Nucleoside Reverse Transcriptase Inhibitor IQP-0410. *Antiviral Res.* **2013**, *100*, 382–391.

(38) Yu, M.; Corsa, A. C.; Xu, S. M.; Peng, B.; Gong, R. Y.; Lee, Y. J.; Chan, K. T.; Mo, H. M.; Delaney, W.; Cheng, G. F. In Vitro Efficacy of Approved and Experimental Antivirals Against Novel Genotype 3 Hepatitis C Virus Subgenomic Replicons. *Antiviral Res.* **2013**, *100*, 439–445.

(39) Peng, H. K.; Chen, W. C.; Lin, Y. T.; Tseng, C. K.; Yang, S. Y.; Tzeng, C. C.; Lee, J. C.; Yang, S. C. Anti-hepatitis C Virus RdRp Activity and Replication of Novel Anilinobenzothiazole Derivatives. *Antiviral Res.* **2013**, *100*, 269–275.

(40) Biswas, S.; Swift, M.; Field, H. J. High Frequency of Spontaneous Helicase-Primase Inhibitor (BAY 57–1293) Drug-Resistant Variants in Certain Laboratory Isolates Of HSV-1. *Antiviral Chem. Chemother.* **2007**, *18*, 13–23.

(41) Ivachtchenko, A. V.; Ivanenkov, Y. A.; Mitkin, O. D.; Yamanushkin, P. M.; Bichko, V. V.; Leneva, I. A.; Borisova, O. V. A Novel Influenza Virus Neuraminidase Inhibitor AV5027. *Antiviral Res.* **2013**, *100*, 698–708.

(42) Kim, M.; Kim, S. Y.; Lee, H. W.; Shin, J. S.; Kim, P.; Jung, Y. S.; Jeong, H. S.; Hyun, J. K.; Lee, C. K. Inhibition of Influenza Virus Internalization by (–)-epigallocatechin-3-gallate. *Antiviral Res.* **2013**, *100*, 460–572.

(43) Martinez-Gil, L.; Alamares-Sapuay, J. G.; Ramana Reddy, M. V.; Goff, P. H.; Premkumar Reddy, E.; Palese, P. A Small Molecule Multi-Kinase Inhibitor Reduces Influenza A Virus Replication by Restricting Viral RNA Synthesis. *Antiviral Res.* **2013**, *100*, 29–37.

(44) Haasbach, E.; Reiling, S. J.; Ehrhardt, C.; Droebner, K.; Ruckle, A.; Hrinicus, E. R.; Leban, J.; Strobl, S.; Vitt, D.; Ludwig, S.; Planz, O. The NF-kappaB Inhibitor SC75741 Protects Mice Against Highly Pathogenic Avian Influenza A Virus. *Antiviral Res.* **2013**, *99*, 336–344.

(45) Sidorov, P.; Gaspar, H.; Marcou, G.; Varnek, A.; Horvath, D. Mappability of Drug-Like Space: Towards a Polypharmacologically Competent Map of Drug-Relevant Compounds. *J. Comput.-Aided Mol. Des.* **2015**, *29*, 1087–1108.

(46) Ruggiu, F.; Marcou, G.; Varnek, A.; Horvath, D. Isida Property-labelled Fragment Descriptors. *Mol. Inf.* **2010**, *29*, 855–868.

(47) Varnek, A.; Fourches, D.; Horvath, D.; Klimchuk, O.; Gaudin, C.; Vayer, P.; Solov'ev, V.; Hoonakker, F.; Tetko, I. v.; Marcou, G. Isida - Platform for Virtual Screening Based on Fragment and Pharmacophoric Descriptors. *Curr. Comput.-Aided Drug Des.* **2008**, *4*, 191–198.

(48) Varnek, A.; Fourches, D.; Solov'ev, V.; Klimchuk, O.; Ouadi, A.; Billard, I. Successful "In Silico" Design of New Efficient Uranyl Binders. *Solvent Extr. Ion Exch.* **2007**, *25*, 433–462.

(49) Evans, B. E.; Rittle, K. E.; Bock, M. G.; DiPardo, R. M.; Freidinger, R. M.; Whitter, W. L.; Lundell, G. F.; Veber, D. F.; Anderson, P. S. Methods for Drug Discovery: Development of Potent, Selective, Orally Effective Cholecystokinin Antagonists. *J. Med. Chem.* **1988**, *31*, 2235–2246.

(50) Kubinyi, H. Privileged Structures and Analogue-Based Drug Discovery. In *Analogue-based Drug Discovery*; Fischer, J. G. R., Ed.; Wiley-VCH Verlag-GmbH & Co. KGaA, Weinheim, Germany, 2006; pp 53–68.

(51) Kireeva, N.; Baskin, I.; Gaspar, H. A.; Horvath, D.; Marcou, G.; Varnek, A. Generative Topographic Mapping (GTM): Universal Tool for Data Visualization, Structure-Activity Modeling and Dataset Comparison. *Mol. Inf.* **2012**, *31*, 301–312.

(52) Gaspar, H. A.; Baskin, I. I.; Marcou, G.; Horvath, D.; Varnek, A. Chemical Data Visualization and Analysis with Incremental Generative Topographic Mapping: Big Data Challenge. *J. Chem. Inf. Model.* **2015**, *55*, 84–94.

(53) Gaspar, H. A.; Baskin, I. I.; Marcou, G.; Horvath, D.; Varnek, A. GTM-Based QSAR Models and Their Applicability Domains. *Mol. Inf.* **2015**, *34*, 348–356.

(54) Gaspar, H.; Marcou, G.; Horvath, D.; Arault, A.; Lozano, S.; Vayer, P.; Varnek, A. Generative Topographic Mapping-Based Classification Models and Their Applicability Domain: Application to the Biopharmaceutics Drug Disposition Classification System (BDDCS). *J. Chem. Inf. Model.* **2013**, *53*, 3318–3325.

(55) Patterson, D. E.; Cramer, R. D.; Ferguson, A. M.; Clark, R. D.; Weinberger, L. E. Neighborhood Behavior: A Useful Concept for Validation of "Molecular Diversity" Descriptors. *J. Med. Chem.* **1996**, *39*, 3049–3059.

(56) Horvath, D.; Jeandenans, C. Neighborhood Behavior of In Silico Structural Spaces with respect to In Vitro Activity Spaces – A Novel Understanding of the Molecular Similarity Principle in the Context of Multiple Receptor Binding Profiles. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 680–690.

(57) Chang, J.-W. A New Cell-Based Clustering Method for High-Dimensional Data Mining Applications. In *Knowledge-Based Intelligent Information and Engineering Systems: Part I of Proceedings of 9th International Conference, KES 2005*, Melbourne, Australia, September 14–16, 2005.

(58) ChemAxon Standardizer, <https://www.chemaxon.com/products/standardizer/> (accessed February, 2009).

(59) Weininger, D. SMILES, A Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules. *J. Chem. Inf. Model.* **1988**, *28*, 31–36.

(60) Hu, Y.; Stumpfe, D.; Bajorath, J. Lessons Learned from Molecular Scaffold Analysis. *J. Chem. Inf. Model.* **2011**, *51*, 1742–1753.

(61) Prado-Prado, F. J.; de la Vega, O. M.; Uriarte, E.; Ubeira, F. M.; Chou, K. C.; Gonzalez-Diaz, H. Unified QSAR Approach to Antimicrobials. 4. Multi-target QSAR Modeling and Comparative Multi-Distance Study of the Giant Components of Antiviral Drug-Drug Complex Networks. *Bioorg. Med. Chem.* **2009**, *17*, 569–575.

(62) Krecmerova, M.; Holy, A.; Piskala, A.; Masojdikova, M.; Andrei, G.; Naesens, L.; Neyts, J.; Balzarini, J.; De Clercq, E.; Snoeck, R. Antiviral Activity of Triazine Analogues of 1-(S)-[3-hydroxy-2-(phosphonomethoxy)propyl]cytosine (Cidofovir) and Related Compounds. *J. Med. Chem.* **2007**, *50*, 1069–1077.