



HAL
open science

QSAR and QSPR modelling of properties of interest for compound screening and safety

Shamkhal Baybekov

► **To cite this version:**

Shamkhal Baybekov. QSAR and QSPR modelling of properties of interest for compound screening and safety. Other. Université de Strasbourg, 2023. English. NNT : 2023STRAF045 . tel-04390291

HAL Id: tel-04390291

<https://theses.hal.science/tel-04390291>

Submitted on 12 Jan 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

ÉCOLE DOCTORALE DES SCIENCES CHIMIQUES

Chimie de la matière complexe – UMR 7140

THÈSE présentée par :
Shamkhal BAYBEKOV

soutenue le : **10 novembre 2023**

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

Discipline / Spécialité : **Chimie**

**Modélisation QSAR et QSPR de
propriétés d'intérêt pour le criblage et la
sécurité des composés**

THÈSE dirigée par :

M. VARNEK Alexandre

M. MARCOU Gilles

Professeur, Université de Strasbourg

Maître des Conférences, Université de Strasbourg

RAPPORTEURS :

Mme. CAMPROUX Anne-Claude

M. LEPAILLEUR Alban

Professeur, Université Paris-Cité

Maître des Conférences, Université de Caen Normandie

AUTRES MEMBRES DU JURY :

M. GALZI Jean-Luc

Mme. CHEDIK Lisa

Directeur de recherches, CNRS UMR 7242

Chargé de recherches, INRS Centre de Lorraine

Abstract

This thesis concerns the development and implementation of chemoinformatics tools to support compound screening campaigns. It covers the following topics: steps for compound selection, assessment of stock solution integrity, quality control of experimental data, development of predictive models, and annotation of screening libraries. The properties of interest include solubility of fragment-like compounds in DMSO, aqueous solubility, skin sensitization, skin permeability, and binding to angiotensin-converting enzyme (ACE2) for the design of biological probes. The publicly available quantitative structure-activity/-property relationship (QSAR/QSPR) models and user-friendly tools developed in KNIME Analytics Platform provide valuable support to researchers without the need for coding expertise. The integration of the developed chemoinformatics tools offers an efficient approach to improving screening outcomes and maximizing efficiency.

Acknowledgment

Personally, I believe that each and every person I encountered and communicated with throughout my life has contributed to the person I have become. Similarly, in QSAR/QSPR modeling, each molecule contributes to establishing structure-activity/-property relationships, however, there are a few that weigh in more than others. In this section, I would like to highlight the "molecules" that influenced me the most and contributed to my Ph.D. project.

Firstly, I would like to express my deep gratitude to Dr. Gilles Marcou and Prof. Alexandre Varnek for sparing a good portion of their time to share their knowledge and nurture my professional skills throughout my Ph.D. Their committed guidance helped me lay a foundation for my development as a chemoinformatician, which will further serve me in my professional growth.

I am grateful to the students whom I had the privilege to consult during their internships: Farah Asgarkhanova, Anna Borisova, Aykhan Israfilli, and Erik Yeghyan. A special note of gratitude goes to Farah for her contribution to the ACE2 inhibition project.

I highly appreciate Dr. Lisa Chedik, Catherine Champmartin, and Dr. Mélanie Mourot (Bousquenaud) for sparing their time to share their expert opinion and knowledge in the domain of skin permeability and sensitization.

I would like to thank Dr. Fanny Bonachera for her technical support in KNIME workflow development and model deployment; Dr. Olga Klimchuk for her help in various aspects of my stay in the lab and in France; Dr. Arkadii Lin, Dr. Yuliana Zabolotna, Dr. Dragos Horvath, Pierre Llompart, Dr. Iuri Casciuc, Dr. Alexey Orlov, Dr. Tagir Akhmetshin, Maxim Shevelev, Regina Pikalyova, Karina Pikalyova, Polina Oleneva, William Bort, Dr. Helena Perez Pena, Sai Prashanth Santhapuri, Dr. Dmitrii Zankov, and Louis Plyer for fruitful discussions and for creating a friendly environment in the lab.

I would also like to express my sincere appreciation to all my teachers from Baku Engineering University for providing me with a solid theoretical background, and specifically to Dr. Yusif Abdullayev for sharing information about the Complex Systems Chemistry (CSC) Master-Ph.D. program, thanks to which I pursued my Master's and Ph.D. studies.

I would like to thank my parents, my grandparents, and my friends for their constant support and being there for me in difficult situations.

I would like to thank the CSC Graduate School funded by the French National Research Agency (CSC-IGS ANR-17-EURE-0016) for a PhD fellowship.

Lastly, I would like to express my sincere gratitude to my wife, Khayala, for her immense support and her belief in me. Her love and care calmed me in the most stressful situations and motivated me to keep moving forward to infinity and beyond.

Contents

1. Résumé en français	1
1.1 Introduction.....	1
1.2 Résultats et discussions.....	3
1.3 Conclusion générale.....	11
1.4 Liste des présentations.....	13
1.5 Liste des publications.....	14
2. Introduction	15
2.1 Screening workflow.....	15
2.2 Chemoinformatics in screening workflow.....	21
2.3 Goal of the thesis.....	22
3. QSAR/QSPR modeling methodology	25
3.1 Molecular standardization.....	25
3.2 Molecular descriptors.....	26
3.3 Machine learning methods.....	26
3.4 Optimization algorithms.....	28
3.5 Evaluation metrics.....	29
3.6 Validation method.....	30
3.7 Applicability domain.....	30
3.8 Consensus modeling.....	31
3.9 Outlier detection.....	31
3.10 ISIDA Predictor software.....	31
3.11 Modelling workflow.....	32
3.12 KNIME workflows.....	32
4. Solubility	35
4.1 Introduction.....	35
4.2 Solubility of fragment-like compounds in DMSO.....	35
4.3 Aqueous solubility.....	47
4.4 Conclusion.....	59
5. Skin-related safety properties	61
5.1 Introduction.....	61
5.2 Skin sensitization.....	62

5.3 Skin permeability.....	82
5.4 Conclusion	106
6. ACE2 selective inhibition.....	107
6.1 Introduction	107
6.2 Data	109
6.3 Methods.....	109
6.4 Results and discussion.....	111
6.5 Conclusion	113
7. Conclusion and perspectives.....	115
7.1 Perspectives	117
8. References	119
9. Appendix	125
10. List of Figures.....	128
11. List of Tables	129

Chapter 1

Résumé en français

1.1 Introduction

Le criblage joue un rôle essentiel dans le processus de découverte et de développement des médicaments.¹ Bien que les essais de criblage accélèrent considérablement ce processus, un grand nombre de résultats négatifs sont générés, ce qui entraîne une consommation inutilement élevée de ressources (humaines et matérielles). L'intégration de méthodes *in silico* dans le processus de criblage¹ permet de biaiser l'ensemble des substances à tester de manière à concentrer les moyens sur les hypothèses les plus fructueuses. Cette thèse vise à développer et mettre en place des outils chémoinformatiques qui accompagneront les campagnes de criblage dans les étapes de collecte et d'analyse des données, de contrôle qualité des données de criblage, de développement de modèles prédictifs spécifiquement adaptés, et d'annotation des bibliothèques de criblage (Figure 1).

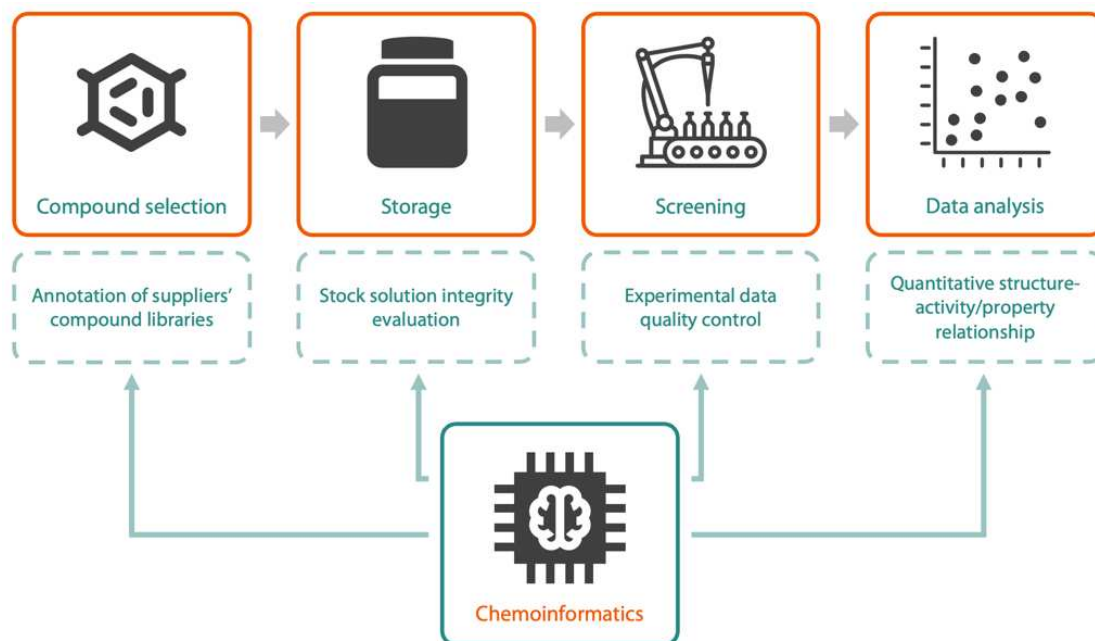


Figure 1. Vue d'ensemble du processus de criblage et de l'application de chémoinformatique à chaque étape du processus.

Les premiers paramètres présentés sont la solubilité dans le DMSO et la solubilité aqueuse. Ils sont d'une grande importance pour le criblage. Le DMSO est un solvant standard pour le stockage des composés organiques ; quand à l'évaluation de la solubilité aqueuse, elle est cruciale lors de la réalisation d'un criblage et, plus tard, pour le développement d'un composé en tête de série. Ensuite, ce travail se concentre sur les essais de criblage pertinents pour la sécurité des produits chimiques : la sensibilisation et la perméabilité cutanée. L'évaluation de la sensibilisation cutanée est aujourd'hui une obligation réglementaire pour l'UE dans le cadre de l'annexe VII de REACH. La perméabilité de la peau est un paramètre crucial, mais difficilement accessible, pour estimer le risque d'un produit chimique. Enfin, la liaison à l'enzyme de conversion de l'angiotensine (ACE2) a été ajoutée aux propriétés modélisées, dans le cadre de la conception de sondes biologiques capables de moduler temporairement l'activité de ACE2 dans différents tissus et organes biologiques.

Ces projets sont le fruit de collaborations avec diverses équipes de recherche et instituts, notamment la Plateforme Intégrée de Criblage de Toulouse (PICT), pour la solubilité dans le DMSO et la solubilité cinétique en milieu aqueux - également avec la Plateforme de Chimie Biologique Intégrative de Strasbourg (PCBIS - UAR 3286) ; l'Institut de Chimie Organique et la société Enamine à Kyiv, en Ukraine, pour l'inhibition sélective de l'ACE2 ; et l'Institut National de Recherche et de Sécurité (INRS) à Nancy, pour la perméabilité et la sensibilisation de la peau. Dans le cadre de cette thèse, des solutions chémoinformatiques ont été développées, notamment des modèles de relations quantitatives structure-activité/-propriété (QSAR/QSPR) accessibles au public, ainsi que

des outils conviviaux pour le déploiement de ces modèles *in silico*. Ces outils sont disponibles sous forme de processus de traitement de données pour l'environnement logiciel KNIME², qui sont faciles à utiliser et ne nécessitent aucune expertise en matière de codage.

1.2 Résultats et discussions

1.2.1 Solubilité des composés apparentés à des fragments dans le DMSO

Le DMSO est un solvant standard largement accepté, utilisé à la fois pour le stockage et le criblage expérimental. Les modèles prédictifs de solubilité dans le DMSO sont utiles pour gérer les collections de substances destinées au criblage, car la saturation des stocks ou des plaques peut passer inaperçue. Cela peut entraîner une estimation incorrecte de la concentration des substances testées et compromettre les résultats des essais biologiques ou des campagnes de criblage. Ce projet se concentre sur le criblage basé sur les fragments (FBS), où l'objectif est de proposer un modèle capable de prédire si un composé peut être concentré à 1 mM dans le DMSO et, donc, être conforme pour une campagne FBS.

La concentration de 1 mM est une concentration nominale typique des échantillons utilisés dans les campagnes FBS.³ Cette valeur a été utilisée comme seuil pour développer un modèle utilisant la structure des fragments pour discriminer ceux qui sont solubles (concentration maximale ≥ 1 mM) ou insolubles (concentration maximale < 1 mM). Le modèle développé a été comparé à un autre modèle disponible publiquement qui utilise un seuil 10 mM pour définir la solubilité (une concentration nominale courante des solutions mères de composés organiques⁴). Cette définition diffère des conditions FBS (Figure 2), mais une comparaison reste possible sous certaines conditions.

Les résultats de cette étude sont résumés dans le Tableau 2. En outre, 34 données erronées ont pu être identifiées au cours de la modélisation. Ces erreurs résultent de problèmes expérimentaux ou de la dégradation des substances en solution. Le modèle développé peut être utilisé pour filtrer et prioriser des composés étiquetés comme "solubles" par le modèle, pour des campagnes de criblage.

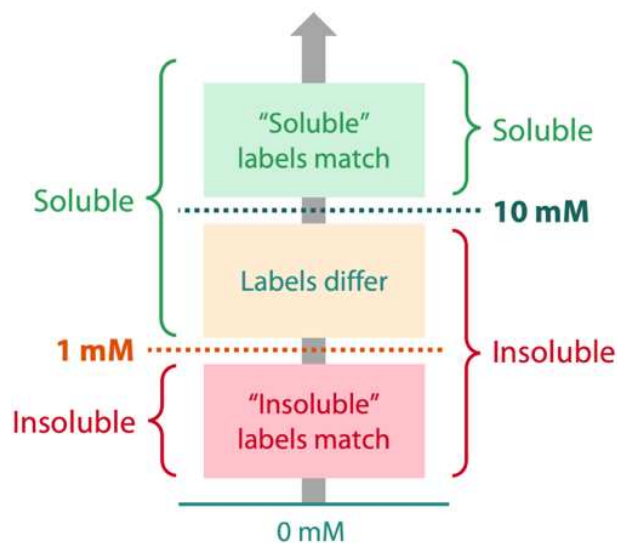


Figure 2. Division de la plage de solubilité en catégories fixes par les seuils pour le criblage basé sur les fragments (1 mM) et pour la formulation de solutions mères (10 mM). Les étiquettes "Soluble" et "Insoluble" coïncident pour les solubilités supérieures à 10 mM et inférieures à 1 mM. Toutefois, dans la plage de 1 à 10 mM, les composés sont considérés comme solubles selon la définition du FBS, mais insolubles selon la définition d'une solution mère.

1.2.2 Solubilité aqueuse

La mesure de la solubilité aqueuse est cruciale dans la découverte et le développement de médicaments, mais son objectif varie aux différents stades de ce processus.⁵ Au début de la découverte de médicaments, l'objectif est d'éliminer rapidement les composés qui ne sont pas suffisamment solubles pour être testés à la concentration maximale d'un essai. La solubilité cinétique est donc privilégiée, car elle peut être mise en œuvre dans une configuration à haut débit, impliquant le criblage d'échantillons préparés à partir de solutions mères.⁵ Aux stades ultérieurs de la découverte et du développement de médicaments, la solubilité est mesurée de manière plus approfondie et tolère un rythme plus lent, pour servir de paramètre à la biodisponibilité et à la sécurité des candidats médicaments. Ces expériences de mesure de la solubilité utilisent une poudre pure comme point de départ et sont appelées essais de solubilité thermodynamique (Figure 3).⁵ Bien que les deux essais soient importants, la solubilité thermodynamique est plus souvent modélisée car elle est considérée comme une quantité thermodynamique, reproductible et ayant une relation directe avec la nature du soluté. Les essais de solubilité cinétique, en revanche, sont moins étudiés car ils sont considérés comme non reproductibles, ne correspondant pas à un équilibre thermodynamique.⁶ L'objectif de ce projet était de développer des modèles qui prédisent spécifiquement la solubilité aqueuse et d'étudier les différences entre solubilité cinétique et solubilité thermodynamique du point de vue de la modélisation.

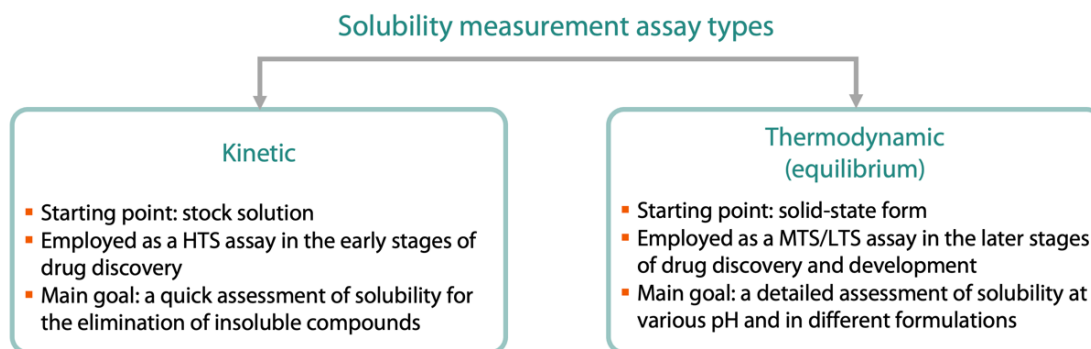


Figure 3. Types d'essais de mesure de la solubilité. HTS, MTS et LTS désignent respectivement un criblage à haut, moyen et bas débit.

L'analyse de plusieurs ensembles de données de solubilité cinétique obtenues par différents protocoles expérimentaux a montré une bonne concordance entre les valeurs mesurées de composés courants, avec des différences inférieures à une unité log (en M). Toutefois, la comparaison des solubilités cinétique et thermodynamique mesurées pour le même ensemble de composés a confirmé l'observation bien établie selon laquelle la solubilité cinétique surestime la solubilité thermodynamique et n'est pas prédictive de celle-ci.

Pour cette étude, le cas du criblage basé sur les fragments a été spécifiquement étudié, et un seuil de classification de 1 mM a été utilisé - le même que pour la modélisation de la solubilité dans le DMSO. Le modèle entraîné sur toutes les données de solubilité cinétique, une fois agrégées et intégrées, est performant sur jeu de données de test (Tableau 2).

Dans le même temps, la solubilité thermodynamique dans l'eau a été réexaminée, en tirant parti de la publication de nouvelles données, abondantes et bien répertoriées. Les modèles publics et nouvellement entraînés de solubilité thermodynamique ont été comparés et leurs performances ont été rationalisées. L'erreur de ces modèles devrait se situer entre 0,8 et 1,0 unité log. Cependant, lorsqu'ils sont appliqués aux données de solubilité cinétique, ils atteignent des performances presque aléatoires (moyenne de la précision balancée (BA) sur l'ensemble de test = 0,56), ce qui montre que la solubilité thermodynamique n'est pas prédictive de la solubilité cinétique.

Les résultats de ce projet mettent en évidence plusieurs points intéressants. Premièrement, les données de solubilité cinétique obtenues à l'aide de différents protocoles de mesure sont en bon accord les unes avec les autres, ce qui indique une bonne reproductibilité interlaboratoire ; simultanément, un protocole complet pour nettoyer les données de solubilité thermodynamique a été défini. Deuxièmement, le mélange des données de solubilité cinétique donne de meilleures performances de modélisation, ce qui suggère que les données de solubilité cinétique sont homogènes et ne dépendent pas de l'essai comme initialement supposé ; d'autre part, les données de solubilité

thermodynamique comprennent souvent des mesures problématiques, par exemple des mesures acquises en dehors de la plage d'utilisation recommandée de la méthode expérimentale employée. Enfin, le modèle formé sur les données de solubilité thermodynamique ne parvient pas à évaluer la solubilité cinétique et *vice versa*, soulignant qu'il s'agit de mesures conceptuellement liées mais différentes. Si l'évaluation expérimentale de la solubilité aqueuse est essentielle, des modèles de solubilité cinétique et thermodynamique sont nécessaires pour soutenir le criblage expérimental et le passage de touches à têtes de série dans développement de médicaments.

1.2.3 Sensibilisation cutanée

La sensibilisation cutanée est une réaction allergique qui se produit lorsque le système immunitaire réagit de manière excessive à une substance particulière qui est entrée en contact avec la peau et l'a pénétrée. Aujourd'hui, de nombreux essais biologiques contrôlent chaque étape de ce processus biologique complexe afin d'évaluer le potentiel de sensibilisation d'une substance chimique.⁷ La voie toxicologique impliquée dans les effets indésirables (AOP) est un cadre conceptuel utilisé pour étudier les phénomènes biologiques, tels que la sensibilisation de la peau, en décomposant des bioprocessus complexes en une série d'étapes appelées événements clefs (KE).⁸ L'AOP de la sensibilisation de la peau se compose de plusieurs KE : liaison des protéines (KE1, ou l'événement initiateur moléculaire (molecular initiating event, MIE)), activation des kératinocytes (KE2), activation des cellules dendritiques (KE3), prolifération des lymphocytes T (KE4) et dermatite de contact allergique (effet indésirable) (Figure 4).⁸ Nos collaborateurs de l'INRS (Institut National de Recherche et de Sécurité) ont développé un nouveau test appelé test des cellules dendritiques de la moelle osseuse (BMDC), qui vise à étudier l'activation des cellules dendritiques (KE3).⁹ L'objectif de ce projet est de contextualiser les données BMDC dans le cadre des bio-essais existants et de construire un modèle QSAR prédisant les résultats de la sensibilisation cutanée sur la base des données expérimentales du test BMDC. La collecte, le traitement des données et la modélisation ont été réalisés avec nos collaborateurs de l'INRS.

Nous avons comparé l'essai BMDC avec des essais *in vitro* et *in chemico* bien connus (DPRA, KeratinoSens™, LuSens, h-CLAT, U-SENS™, mMUSST) sur la base d'un ensemble de composés communs qui ont été testés expérimentalement à l'aide d'essais sélectionnés. La comparaison a été faite par rapport à un test *in vivo* bien établi : l'essai de stimulation locale des ganglions lymphatiques (LLNA). Les résultats ont montré que l'essai BMDC était légèrement plus performant (BA = 0,86) que les autres essais (BA ≤ 0,84). Le modèle QSAR construit sur les données BMDC a montré de bonnes performances (Tableau 2) démontrant la cohérence interne de l'ensemble de données.

En résumé, les résultats de l'essai BMDC ont montré une meilleure performance par rapport à d'autres essais de sensibilisation cutanée *in vitro* et *in chemico* en comparaison à

l'essai *in vivo* LLNA choisit comme référence. Le modèle QSAR développé sur la base des données BMDC peut être utilisé pour accompagner l'évaluation expérimentale préliminaire de la sensibilisation cutanée des composés d'intérêt et pour guider les experts dans la hiérarchisation des essais de certains composés par rapport à d'autres.

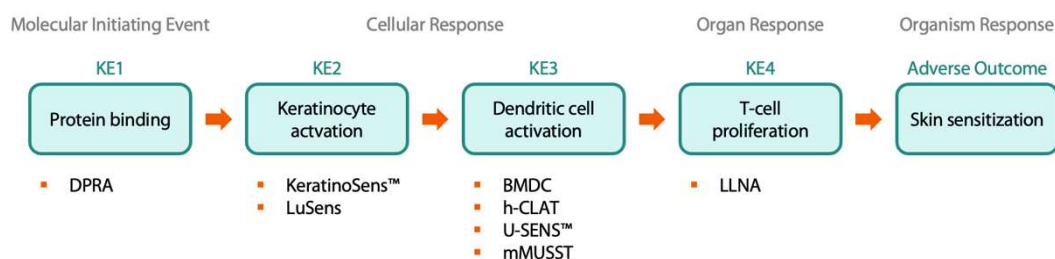


Figure 4. Voie d'expression des effets indésirables (AOP) du processus de sensibilisation cutanée. Une liste non exhaustive d'essais bien connus décrivant chaque événement clef (KE) est donnée sous leur KE respectif.

1.2.4 Perméabilité cutanée

L'évaluation du taux de perméation cutanée est importante non seulement pour les produits chimiques pharmaceutiques et cosmétiques, mais aussi pour les substances toxiques industrielles, car les travailleurs peuvent y être exposés lors de manipulations.¹⁰ Les expériences de perméabilité cutanée prennent souvent beaucoup de temps et dépendent de la disponibilité d'échantillons de peau fraîche. Bien que la validation expérimentale soit nécessaire, les méthodes *in silico* développées pour l'évaluation virtuelle de l'absorption cutanée peuvent être utilisées pour une évaluation préliminaire.¹¹ Le développement de tout modèle QSPR implique l'utilisation de sources de données pour l'entraînement, qui doivent de préférence contenir des chimiotypes variés et être à jour. HuskinDB est la plus grande base de données connue sur la perméabilité cutanée, mais elle est constituée de données provenant de sources bibliographiques publiées jusqu'en 2011 et n'a pas été mise à jour depuis.¹² Pour résoudre ce problème, l'une de nos tâches a été de compiler un nouvel ensemble de données sur la perméabilité cutanée publiées entre 2012 et 2021, afin de compléter HuskinDB. Ce nouveau jeu de données est prêt et publié. Il est utilisé pour entraîner de nouveaux modèles QSPR couvrant un espace chimique plus large. La collecte, le traitement des données et la modélisation ont été réalisés avec nos collaborateurs de l'INRS.

Les nouvelles données ont été collectées à partir d'articles publiés entre 2012 et 2021 et leur pertinence pour les mesures de perméabilité cutanée a été évaluée (Figure 5). Une fois la sélection des articles pertinents terminée, le coefficient de perméabilité cutanée et d'autres métadonnées ont été extraits des documents. Un filtrage supplémentaire des données a consisté à supprimer les substances de composition inconnue ou variable (UVCB) et à conserver les points de données obtenus à l'aide de dispositifs expérimentaux définis. En conséquence, 202 nouveaux points de données (110 composés) ont été extraits

de 621 articles, et le nouvel ensemble de données, SkinPiX, a été mis à la disposition du public¹³.

Le modèle QSPR formé sur la combinaison de la base de données HuskinDB et de la base de données SkinPiX nouvellement compilée est plus performant ($RMSE_{5-CV} = 0,7$) (Tableau 2) que le modèle développé uniquement sur la base de données HuskinDB ($RMSE_{5-CV} = 0,76$). L'ensemble d'entraînement plus important du nouveau modèle QSPR (195 composés par rapport aux 123 composés de l'ensemble d'entraînement du modèle HuskinDB) permet de couvrir une plus grande partie de l'espace chimique.

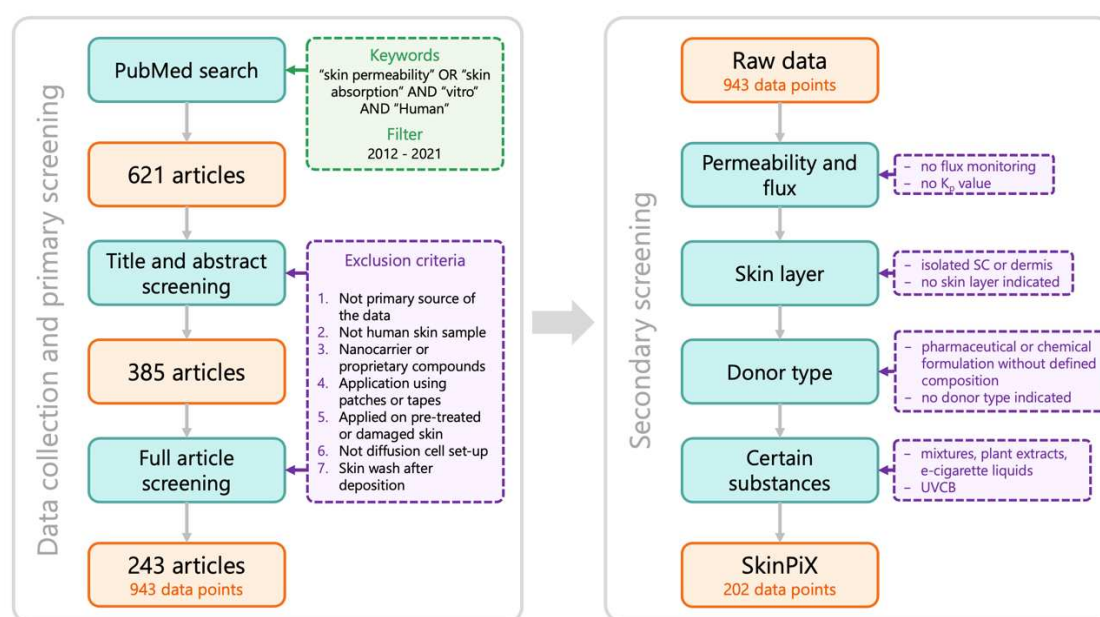


Figure 5. Processus de collecte et de filtrage des données de la nouvelle base de données SkinPiX. "SC" signifie stratum corneum. Le processus suit deux étapes principales. Tout d'abord, les publications scientifiques pertinentes ont été recherchées dans PubMed. Ensuite, les données sur la perméabilité de la peau ont été extraites avec leurs métadonnées. Seules les données répondant à des critères spécifiques ont été conservées, comme illustré.

1.2.5 Inhibition sélective de l'ACE2

L'enzyme de conversion de l'angiotensine 2 (ACE2) est une enzyme liée à la membrane cellulaire de nombreux types de cellule : pulmonaires, cardiaques et rénales en particulier. L'ACE2 fait partie du système rénine-angiotensine-aldostérone (RAAS), qui régule la pression artérielle et l'équilibre des fluides dans l'organisme.¹⁴ Outre son rôle dans le RAAS, l'ACE2 est utilisée par les coronavirus (y compris le SARS-CoV-2) comme porte d'entrée pour infecter une cellule et accéder à sa machinerie.¹⁵ D'où l'intérêt de découvrir des sondes biologiques pouvant être utilisées pour moduler l'activité et comprendre le rôle biologique de l'ACE2. Au cours de ce projet, le criblage virtuel de la collection de composés en stock de la société Enamine (2,6 millions de composés) et d'un ensemble de 4080 composés

précédemment conçus *in silico* a été réalisé. Les résultats ont permis d'établir une liste de molécules susceptibles de se lier sélectivement à l'ACE2. Les méthodes de criblage virtuel appliquées dans le cadre de ce projet comprenaient des méthodes d'amarrage moléculaire (*docking*), de pharmacophore et de modélisation QSAR (Figure 6).

Trois modèles de classification QSAR ont été développés pour prédire l'inhibition des enzymes ACE2, ACE et NEP (Tableau 2). L'objectif était de trouver des molécules susceptibles de mettre en évidence l'effet de l'ACE2 par rapport à l'ACE ; comme la NEP régule la durée de vie des peptides natriurétiques, il peut être gênant pour observer sélectivement les ACE / ACE2. Par conséquent, les modèles QSAR de l'ACE et de la NEP ont été utilisés pour identifier les liants sélectifs potentiels de l'ACE2. Finalement, 63 inhibiteurs sélectifs potentiels de l'ACE2 ont été identifiés dans la collection de composés en stock d'Enamine et dans l'ensemble des composés conçus *in silico*. Des essais expérimentaux sont en cours en Ukraine.

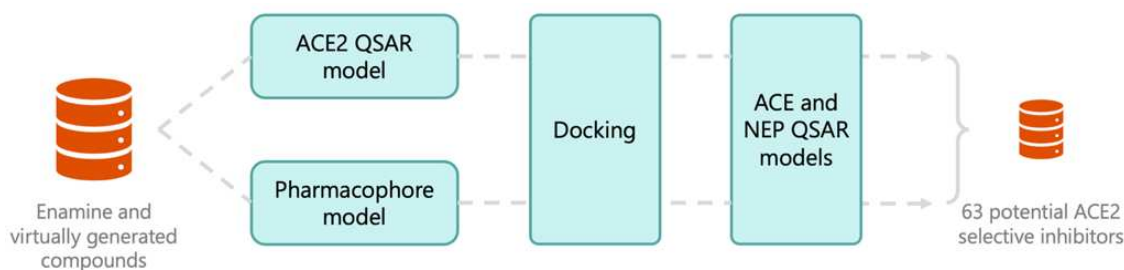


Figure 6. Aperçu des étapes du criblage virtuel pour identifier les inhibiteurs sélectifs de l'ACE2.

1.2.6 Procédures automatisées avec KNIME

Certains des modèles développés dans cette thèse ont été générés à l'aide de procédures automatisées dans la plateforme KNIME². Ces procédures KNIME incluent l'ensemble des étapes de modélisation : la standardisation des structures des molécules, la préparation des jeux de données pour la validation croisée en k paquets, le calcul des descripteurs moléculaires à l'aide du logiciel ISIDA Fragmentor, l'entraînement et la validation de modèles SVM, la préparation du modèle consensus et l'intégration dans le logiciel ISIDA Predictor (qui permet d'utiliser le modèle consensus et intègre le domaine d'applicabilité), l'application du modèle à de nouvelles structures chimiques (Tableau 1). L'utilisation des procédures KNIME ne nécessite aucune connaissance en matière de codage et est donc conviviale et facile à comprendre. Ils ont été développés pour les systèmes d'exploitation Linux et Windows. Les procédures KNIME sont disponibles sur demande à l'aide d'un formulaire sur le site web du Laboratoire de Chémoinformatique (https://infochim.chimie.unistra.fr/?page_id=11). Les procédures sont illustrées dans les figures en annexe.

Tableau 1. La liste des procédures automatisées KNIME développées.

Nom	Description
1_standardization	Standardisation des structures moléculaires
2_ExtCV_data_partitioning	Division de l'ensemble de données en un nombre défini d'ensembles d'entraînement et de test pour la validation croisée externe
3_ExtCV_descriptor_calculation	Calcul des descripteurs ISIDA
4_ExtCV_modeling_CLS (ou _REG)	Entraînement des modèles de classification (ou de régression) intégré à une validation croisée externe
5_ExtCV_consensus_preparation_CLS (ou _REG)	Prépare un modèle consensus de classification (ou de régression) pour chaque paquet de validation croisée externe.
6_ExtCV_application	Applique un modèle consensus pour l'ensemble de test correspondant à un jeu d'entraînement au cours de la validation croisée externe.
7_ExtCV_evaluation	Évalue la performance prédictive de la validation croisée externe
8_final_consensus_preparation_CLS (ou _REG)	Effectue toutes les étapes de la modélisation pour préparer le modèle consensus final

1.3 Conclusion générale

Tableau 2. La liste des modèles QSAR/QSPR développés, la taille de leurs ensembles d'apprentissage et les valeurs de performance. VC 5-fois - la validation croisée 5-fois ; BA - la précision balancée ; RMSE - racine de l'erreur quadratique moyenne. Remarque concernant la BA : BA = 0,5 - prédiction aléatoire ; BA = 1 - prédiction parfaite. Remarque concernant le RMSE : plus le RMSE est petit, meilleur est le modèle. Entre parenthèses, la taille de l'ensemble de test est indiquée.

Propriété / Activité	Taille de l'ensemble d'entraînement	Méthode de validation	Performance (BA)
Solubilité dans le DMSO	788	VC 5-fois	0.78
Solubilité cinétique aqueuse	56132	Ensemble de test (17666)	0.84
Sensibilisation cutanée	117	VC 5-fois	0.82
Inhibition de l'ACE2	668	VC 5-fois	0.97
Inhibition de l'ACE	591	VC 5-fois	0.83
Inhibition de la NEP	464	VC 5-fois	0.79
Propriété / Activité	Taille de l'ensemble d'entraînement	Méthode de validation	Performance (RMSE)
Perméabilité cutanée	195	VC 5-fois	0.7
Solubilité thermodynamique aqueuse*	42159	Ensemble de test (5728)	0.59

* Le modèle QSPR a été développé par mon collègue doctorant Pierre Llompart.

En plus des 8 modèles QSAR/QSPR développés (Tableau 2), les procédures automatisées utilisées dans leur développement permettent à un utilisateur de générer rapidement ses propres modèles pour la propriété qui l'intéresse. Tous les modèles QSAR/QSPR développés au cours de cette thèse sont accessibles au public sur le serveur web du Laboratoire de Chémoinformatique (<https://chematlas.chimie.unistra.fr/cgi-bin/predictor2.cgi>). Ces modèles et l'ensemble des données de haute qualité produites sont utiles pour annoter les collections de composés pour les paramètres importants couverts dans cette thèse. D'autres travaux devraient porter sur d'autres paramètres importants pour la plate-forme de criblage, tels que la cytotoxicité, la cancérogénicité et d'autres paramètres importants pour mesurer le risque des produits chimiques, par exemple la bronchiosorption.

Les modèles QSAR/QSPR développés sont accessibles sur le service web (<https://chematlas.chimie.unistra.fr/cgi-bin/predictor2.cgi>) en sélectionnant d'abord "General kind of property" et ensuite "Property to model" (Tableau 3).

Tableau 3. La liste des modèles développés et les moyens d'y accéder. Tous les modèles (sauf la solubilité aqueuse thermodynamique) sont accessibles sur la page web <https://chematlas.chimie.unistra.fr/cgi-bin/predictor2.cgi>.

Propriété / Activité	"General kind of property"	"Property to model"
Solubilité dans le DMSO	PhysProp	Solubility_DMSO_2CIs
Solubilité cinétique aqueuse	PhysProp	Kinetic_solubility_2CIs
Perméabilité cutanée	PhysProp	Skin_permeability_Reg
Sensibilisation cutanée	Activity	Skin_sensitization_BMDC_2CIs
Inhibition de l'ACE2	Activity	ACE2_2CIs
Inhibition de l'ACE	Activity	ACE_2CIs
Inhibition de la NEP	Activity	NEP_2CIs
Solubilité thermodynamique aqueuse*	-	-

* Le modèle QSPR a été développé par mon collègue doctorant Pierre Llompart. Le modèle est disponible sur une page web séparée : https://chematlas.chimie.unistra.fr/WebTools/predictor_solubility.php

1.4 Liste des présentations

Conférences nationales et internationales

1. **Baybekov S.**, Marcou G., Ramos P., Saurel O., Galzi J.-L.; Varnek A.; Prediction of DMSO solubility for fragment-based screening; 8th Chemoinformatics Strasbourg Summer School; Strasbourg, France; 27 juin 2022 – 1 juillet 2022; Affiche
2. Asgarkhanova F., **Baybekov S.**, Horvath D., Zabolotna Y., Marcou G., Varnek A.; Computer-aided design of selective chemical probes of angiotensin-converting enzyme 2; 8th Chemoinformatics Strasbourg Summer School; Strasbourg, France; 27 juin 2022 – 1 juillet 2022; Affiche
3. **Baybekov S.**, Marcou G., Ramos P., Saurel O., Galzi J.-L.; Varnek A.; Prediction of DMSO solubility for fragment-based screening; 9th French-Japanese Workshop on Computational Methods in Chemistry; Strasbourg, France; 24 avril 2023 – 25 avril 2023; Affiche
4. Llompart P., Minoletti C., **Baybekov S.**, Horvath D., Marcou G., Varnek A.; Will we ever be able to accurately predict solubility?; 9th French-Japanese Workshop on Computational Methods in Chemistry; Strasbourg, France; 24 avril 2023 – 25 avril 2023; Affiche

Autres présentations (journée des doctorants, journées d'UMR, etc.)

1. **Baybekov S.**; Prediction of DMSO solubility for fragment-based screening; Journée Scientifique de l'UMR 7140; Strasbourg, France; 10 mai 2022; Oral

1.5 Liste des publications

1. **Baybekov, S.**; Marcou, G.; Ramos, P.; Saurel, O.; Galzi, J.-L.; Varnek, A. DMSO Solubility Assessment for Fragment-Based Screening. *Molecules* **2021**, *26* (13), 3950. <https://doi.org/10.3390/molecules26133950>.
2. Chedik, L.; **Baybekov, S.**; Cosnier, F.; Marcou, G.; Varnek, A.; Champmartin, C. SkinPiX : An Update of Skin Permeability Data based on A Systematic Review of Recent Research. *Scientific Data* (submitted).
 - a. Jeu de données : Chedik, L.; Baybekov, S.; Cosnier, F.; Marcou, G.; Varnek, A.; Champmartin, C. Données de Réplication Pour : SkinPiX (Skin Permeation of Identified Xenobiotics): An Update of Skin Permeability Data Based on A Systematic Review of Recent Research, 2023. <https://doi.org/10.57745/7FHQOY>.
3. **Baybekov, S.**; Llompart, P.; Marcou, G.; Gizzi, P.; Galzi, J.-L.; Ramos, P.; Saurel, O.; Bourban, C.; Minoletti, C.; Varnek, A. Kinetic solubility: experimental and machine-learning modeling perspectives. *Molecular Informatics* (submitted).
 - a. Jeu de données : Baybekov, S.; Llompart, P.; Marcou, G.; Gizzi, P.; Galzi, J.-L.; Ramos, P.; Saurel, O.; Bourban, C.; Minoletti, C.; Varnek, A. Données de Réplication Pour : Kinetic Solubility: Experimental and Machine-Learning Modeling Perspectives, 2023. <https://doi.org/10.57745/ZWS0WC>.
4. Llompart, P.; Minoletti, C.; **Baybekov, S.**; Horvath, D.; Marcou, G.; Varnek, A. Will we ever be able to accurately predict solubility? *Scientific Data* (submitted).
 - a. Jeu de données : Llompart, P.; Minoletti, C.; Baybekov, S.; Horvath, D.; Marcou, G.; Varnek, A. Données de Réplication Pour : Will we ever be able to accurately predict solubility?, 2023. <https://doi.org/10.57745/CZVZIA>.
5. Chedik, L. and **Baybekov, S.**; Marcou, G.; Cosnier, F.; Mourot-Bousquenaud, M.; Jacquenet, S.; Varnek, A.; Battais, F. Benchmarking and QSAR Modeling of BMDC Assay for Identifying Sensitizing Chemicals. *Regulatory Toxicology and Pharmacology* (submitted).
 - a. Jeu de données : Chedik, L. and Baybekov, S.; Marcou, G.; Cosnier, F.; Mourot-Bousquenaud, M.; Jacquenet, S.; Varnek, A.; Battais, F. Données de Réplication Pour : Benchmarking and QSAR Modeling of BMDC Assay for Identifying Sensitizing Chemicals, 2023. <https://doi.org/10.57745/PPAMKY>

Chapter 2

Introduction

2.1 Screening workflow

Drug discovery and development pathway involves exploration and exploitation of chemical space to find the most effective drug candidates. While attempts to design analogs of known bioactive compounds can yield positive results, this approach is limited to investigation of a small number of similar compounds representing limited regions of chemical space. In addition, the industrial development of a new drug must take into account existing patents.¹⁶ A systematic screening campaign can alleviate these issues, by testing hundreds to millions of structurally diverse compounds which cover larger chemical space. This is implemented as a high-throughput screening (HTS), which typically involves testing tens of thousands of compounds per hour.¹⁷ Thus, HTS campaigns allow quick profiling of compound collections for physicochemical properties and/or bioactivities. The most promising molecules, the *hits* of the screening campaign, are then used as starting points to be optimized into *lead* compounds and later to *drug candidates*.

2.1.1 Types of screening campaigns

Screening campaigns can be categorized based on factors, such as the rate of measurement, the involved assay type, screening approach (Table 4).¹⁸ Generally, the number of analyzed compounds per measurement campaign decreases through the drug discovery and development pipeline, while focusing on fewer but more promising lead molecules.¹⁹ The screening rate also varies depending on the chosen screening approach and the applied assay.^{19,20} Here, “classical” screening approach refers to screening of a compound library composed of common compounds of varying molecular weight and chemotypes using automatized platforms.

Table 4. Screening campaign types categorized by measurement rate, approach, and assay type.

Screening rate	Screening approach	Assay type
<ul style="list-style-type: none"> ▪ High-throughput ▪ Medium-throughput ▪ Low-throughput 	<ul style="list-style-type: none"> ▪ "Classical" screening ▪ Fragment-based screening ▪ DNA-encoded library screening ▪ Virtual screening 	<ul style="list-style-type: none"> ▪ Bioassays <ul style="list-style-type: none"> - Target-based screening - Phenotypic screening ▪ ADMET profiling ▪ Physicochemical profiling

The concept of "fragment-based" screening (FBS) is based on the identification of small molecules, called fragments, that fit efficiently to a binding site of the target protein, based on the same technological platforms as for "classical" screening.^{21,22} There are several definitions of what is a fragment²²⁻²⁴, but they all cover the idea of a small molecular weight compound with a limited number of chemical functions. The potency of a single fragment is usually low: binding affinity in a μM -mM range. However, these fragments are convenient platforms to optimize both their biological activities and other desirable properties, such as their solubility in water. This approach allows screening much smaller chemical libraries (500 - 10000 molecules²⁵) compared to HTS libraries (hundreds of thousands to millions of molecules^{21,25}). Despite the smaller size of fragment libraries, their structural diversity and the diversity of the chemistry that can be implemented on them, allows for a better chemical space coverage for the same size of HTS library.²¹ Schuffenhauer et al.²⁶ also reported that hit rates of FBS campaigns were 10-1000 times higher than for HTS campaigns. In summary, FBS approach is a good alternative to the conventional HTS approach, although it has different planning and logistic constraints.

An emerging approach is to test a complex mixture of compounds in a one-pot experiment and use a powerful DNA amplification technique to deconvolute the signal. The "DNA-encoded library" (DEL) screening campaigns allow testing millions to billions of compounds against a biological target at once.²⁷ The procedure includes tagging each molecule with an identifying DNA, incubating the library of DNA-tagged compounds in a mixture with a target protein, washing away non-binding ligands, and identification of the binder by DNA sequencing.²⁸ The DEL screening approach is cost-, time- and material resources-efficient. Nevertheless, DEL preparation is restricted to soft chemical synthesis conditions in order to preserve the integrity of the DNA tags.²⁹

Virtual screening (VS) is an *in silico* approach that is used as a filter to cherry pick compounds possessing desired property profile. VS can be divided into two broad categories: structure-based virtual screening (SBVS) and ligand-based virtual screening (LBVS).³⁰ SBVS concept is based on a protein structure and it involves scoring of a molecule's fit to the binding sites of the target using molecular docking and/or structure-based pharmacophore modelling methods. LBVS utilizes a dataset of known actives and inactives to build quantitative structure-activity relationship (QSAR) and ligand-based pharmacophore models that will then be applied to a new dataset to identify candidate molecules. Both SBVS and LBVS methods can be used individually or in consensus to rank molecules based on their activity against the target protein. Virtual screening is often used

in combination with experimental screening methods to prioritize testing of certain molecules, hence, saving time and resources.

Although each of the screening approaches has its own advantages and limitations, a recent study³¹ showed the preference of certain approaches over the others (Figure 7). The author scrutinized 156 clinical candidates published in the Journal of Medicinal Chemistry between 2018 and 2021 to identify most commonly used lead generation strategies that yielded drug candidates. The results show that the main strategy employed to generate a lead molecule is based from the hits identified from the previous studies (59%). The next approaches are related to HTS involving random (21%) and directed (11%) screenings. Origins of the remaining clinical candidates are distributed among FBS (7%), VS (1%) and DEL screening (<1%). This study shows that despite the emergence of new approaches such as fragment-based screening and DNA-encoded library screening, HTS and derivation from hits identified from previous campaigns still remain as favorites for generation of lead molecules.

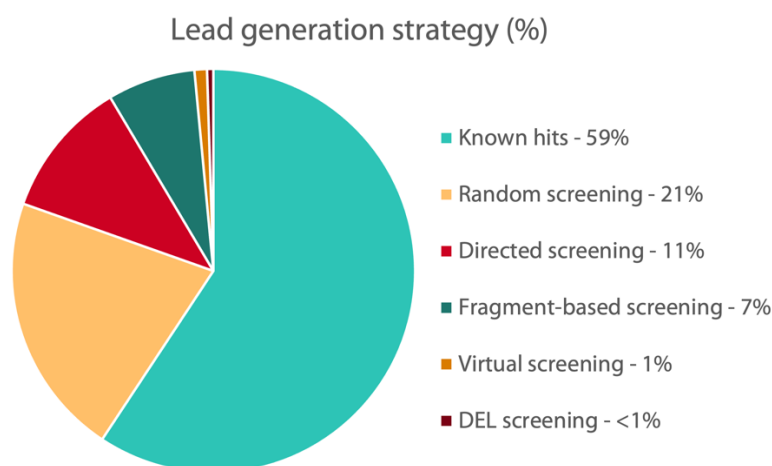


Figure 7. Distribution of lead generation strategies used in 156 successful hit-to-clinical campaigns. The figure is adapted from Brown³¹.

Screening of molecules is also performed to assess their bioactivity, ADMET and physicochemical properties. The solubility and lipophilicity are often assessed upstream to bioassays in order to check the compliance of compounds with the constraints of the assays.¹⁹ Screening and optimization of absorption, distribution, metabolism, excretion and toxicity (ADMET) properties are investigated during both drug discovery and development stages to ensure drug's bioavailability and safety.³²⁻³⁴ Based on the scope of the screening process, bioassays can be generally differentiated into two categories, namely target-based and phenotypic screening types.³⁵ The former is a molecular approach which focuses on the interaction of test molecules with a defined biological target, such as a protein. Examples of target-based assays include screens measuring enzyme inhibition, receptor binding, protein-protein interaction. Unlike target-based screening, phenotypic screening is an empirical approach focusing on phenotypic change. For this reason, they require a complete biological entity to work, such as a cell line. On the other hand, they require no prior knowledge about the identity of the target and can actually help to identify such

target. Examples of phenotypic assays include cell viability tests, changes in the expression of proteins. While both assay categories are complementary approaches, typically, phenotypic screening yields “first-in-class” drugs, whereas target-based screening results in “best-in-class” drug molecules due to availability of the target’s structural information.³⁵

2.1.2 Steps of screening workflow

Despite the variety of existing screening types and approaches, the workflow of an experimental screening campaign can be generalized into several steps in the process (Figure 8): (1) chemical library design, (2) stock preparation, (3) sample preparation, (4) performing the test, (5) data acquisition, (6) data analysis.

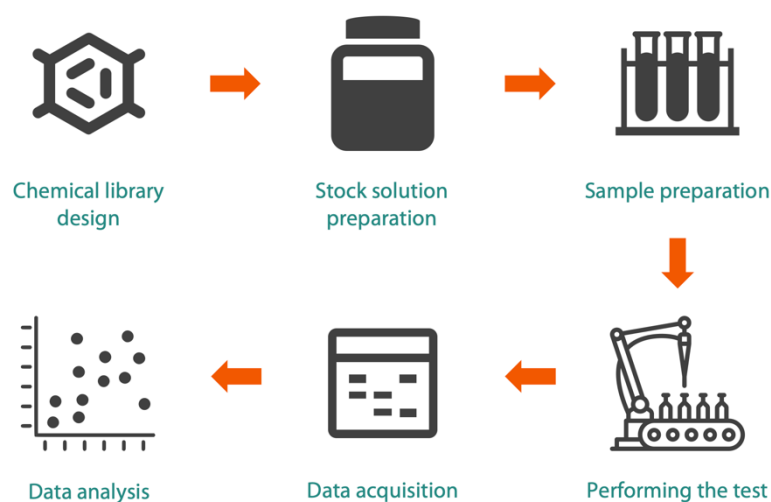


Figure 8. Simplified overview of a screening campaign: (1) chemical library design, (2) stock preparation, (3) sample preparation, (4) performing the test, (5) data acquisition, (6) data analysis.

Screening libraries are provided commercially by specialized synthetic companies^{36,37} or by suppliers who assembled their libraries using compounds synthesized in academic laboratories^{38,39}. A variety of general and focused screening libraries have been compiled to suit every screening campaign depending on its nature.⁴⁰ The selected compounds are commonly distributed in either pure form or stock solution.³⁹ Dimethyl sulfoxide (DMSO) is typically chosen as a solvent for stock solution due to its high solubilizing ability.¹⁸ Concentration of the stock solution can vary depending on the nature of screening library: for instance, for general library a commonly used concentration is 10 mM³⁹, whereas for fragment libraries 50-100 mM stock solutions are preferred²⁴. Stock solutions are usually stored frozen either at 4°C or -20°C.³⁹ Sample preparation and screening steps are assay-dependent and can vary significantly. Automated screening campaigns utilize detection modalities like absorbance, fluorescence, luminescence, radiometry.¹⁸ The data acquired during the measurement are collected by specific software and can be used for further analysis. Depending on the goal, screening campaigns can be performed iteratively to eliminate undesirable compounds, while adding analogues of active molecules.

2.1.3 Expenditures of screening campaigns

Preparation and conduct of screening require a lot of time, material, and human resources. This section provides information about cost of screening library preparation and performing HTS campaign, followed by specific examples. It is worth to mention that, due to variety of available screening set-ups it is impossible to derive one estimate for cost and time of screening campaigns. Therefore, all values provided in this section are approximate and subject to specific cases and screening set-ups.

Goodnow⁴¹ reported that on average the cost estimate for preparation of a one-million-compound collection for use in HTS campaigns can vary from 50 million to 5 billion US\$, depending on the commercial availability and the amount of required compounds (Table 5). He also reported the approximate cost of performing a HTS campaign for a one-million-compound library would be in the range from 100'000 to 200'000 US\$ based on the estimate of 0.07-0.2 US\$ per well.

Table 5. Cost estimates per sample (US\$) for preparation of compounds for HTS campaigns. The table is adapted from Goodnow⁴¹.

Cost estimates	Commercially available compound (10 mg)	Custom synthesis (mg to g)	Purified compound from parallel or combinatorial methods (10 mg)
Lower	50	1000	100
Higher	250	5000	500

Another study by Burbaum⁴², where he analyzed the effect of miniaturization of HTS campaigns, showed that the cost of testing a set of 106 compounds against 40 targets would cost 35 million US\$ if performed in 96-well plates and 1.1 million US\$ in 1536-well plates. These numbers include the cost of compounds, bioreagents, plates and other expenses. From these estimates one could derive costs of one compound screened against one target to be 8000 US\$ (performed in 96-well plates) and 260 US\$ (performed in 1536-well plates).

It is also reported¹⁸ that in 2014 for 10 large pharma companies the mean annual capital budget dedicated for screening was set to 3.5 million US\$ and 3.9 million US\$ for reagents and consumables. These values represent a fraction of the 13 pharmaceutical companies' overall research and development (R&D) spending, which over the eight-year period from 2006 to 2014 ranged from 22 billion to 72 billion US\$. Based on the number of new molecular entities (NMEs) registered during the same time period, these companies had an R&D efficiency of 3 to 32 billion US\$ per NME. This emphasizes the need to take costs into account during all phases of drug discovery innovation, especially while conducting screening.

However, these activities are not the exclusivity of private interests and screening campaigns are efficiently implemented by public facilities. For instance, the Chimiothèque Nationale Française has been initiated in 2000⁴³ and since then provides chemical libraries from French academic libraries to screening platforms at cost price. The know-how for

resynthesis of hits and the hit-to-lead development is matter of fair sharing of intellectual property between synthetic chemistry and drug discovery teams. This facility is completed by a network of experimental and computational laboratories able to implement cutting edge bioassays technologies and in silico methods. This platform is called ChemBioFrance⁴⁴ and is answering free of charge to technological screening questions. The aim of such facility is to provide research groups with an affordable access to these technologies. Thus, a user has access to it at reduced cost, after examination of his/her scientific question.

Other initiatives have focused on providing users with experimentally annotated chemical libraries: the MLSMR (Molecular Libraries Small Molecule Repository), now available in the NExT screening library⁴⁵, or the EU-OPENSREEN (European Infrastructure of Open Screening Platforms for Chemical Biology)⁴⁶.

Bioassays

One of our industrial collaborators reported the average cost of a screening data point to be approximately 1 euro, taking into account the expense associated with full-time equivalent employees (FTEs). Although the cost of reagents for HTS cell-based assay can vary from 0.05 euros to 5 euros per sample, the recommended cost is 0.3 euros. For instance, a cost of an acetylcholinesterase assay is estimated to be less than 0.1 euros per data point. The measurement cost of other endpoints, such as apparent permeability is 90 euros per sample and 50 euros per sample for inhibition of CYP3A. It is important to note that the prices of the latter two cases is for testing outside HTS, and the final cost of the measurement when integrated in HTS is much higher, considering the analysis of several concentrations and usage of expensive analytical systems, such as mass spectrometry.

Skin permeability

The details about skin permeability measurement were provided by INRS (Institut National de Recherche et de Sécurité). The experimental protocol employed at INRS involves usage of radiolabeled compounds, which can cost 10'000 – 35'000 euros per 250 μ Ci of one compound (curie (Ci) is a unit of quantity of radioactive atoms). Costs of other consumables such as pipette tips, solvents, etc., were not reported. The other expenses involve equipment and analytical instrument purchases, such as an automatic sampler (35'000 euros), a radioactivity counter (50'000 euros) and Franz cells (200 euros per 5.5 mL Franz cell), a specific glassware used in skin permeability experiments. The duration of a measurement campaign for 1-2 new substances is distributed over tasks as follows: 1 week for development of method for chemical sample preparation; 8 weeks for preliminary experiments and tuning experimental set-up; a half day for preparation of chemical sample for experiment; 2 hours for setting up the experiment; 20-40 hours for experiment; 1 week for quantification, analysis of results and cleaning. Since, the experiment is performed 5 times, and considering the fact that some of the operations are performed in parallel, in total 18 weeks (8 weeks of preliminary experiments and 10 weeks for 5 experiments) are

required to obtain skin permeability values of 1-2 new substances. Three to four researchers are necessary for execution of all steps.

Skin sensitization

Collaborators from INRS have also shared details of skin sensitization assay, namely, bone marrow-derived dendritic cell (BMDC) assay. The provided costs are related to the recent measurement campaign of one family of compounds composed of 22 representative compounds: 13 mice (500 euros) to obtain 38 bone marrows; 22 compounds (2000 euros); cell culture consumables, such petri dishes, media and antibodies for flow cytometry (8200 euros). In total, the cost of this campaign was about 10'700 euros (\approx 500 euros per substance). The total time spent on one experiment is about 2 months and can be described as follows: 1 month for checking of substances' feasibility (not toxic against dendritic cells); 10 days for one run, which is repeated 3 times (\approx 1 month). Since maximum 4 substances can be checked during one experiment, the analysis of 22 chemicals would require approximately 5 runs, that would take about a year to complete. In general, 2 researchers are required to carry out the whole measurement campaign.

2.2 Chemoinformatics in screening workflow

As mentioned in "Types of screening campaigns" section, virtual screening (VS) is commonly applied prior to or in parallel with experimental screens to aid in selection and prioritization of virtually determined actives. Although, virtual annotation of screening libraries is important, chemoinformatics offers more application cases in screening domain (Figure 9). For instance, an approach to design a diversity oriented screening library is to start with a clustering of molecules in the catalog of suppliers followed by cherry-picking chemical structures homogeneously across the clusters.⁴⁷ This approach is also applied during the selection of candidate compounds for a secondary screening if too many actives have been identified during a primary screening.¹ Another possible application of chemoinformatics is evaluation of the stock solutions' integrity. Degradation signs can be determined by comparing experimentally determined solubility of compounds in DMSO stock solutions with predicted solubility values.⁴⁸ This is achieved by building a quantitative structure-property relationship (QSPR) model trained to predict solubility of compounds in DMSO. The same approach is used to assess the experimental data quality, by fitting a QSPR/QSAR model to screening data and applying the model on the same data.⁴⁹ The outlying data points are then identified and examined. Later the high-quality experimental data obtained from screening campaigns can be used alone or in combination with public data sources to train in-house predictive models, which can be used in future to predict properties or activities of new molecules.

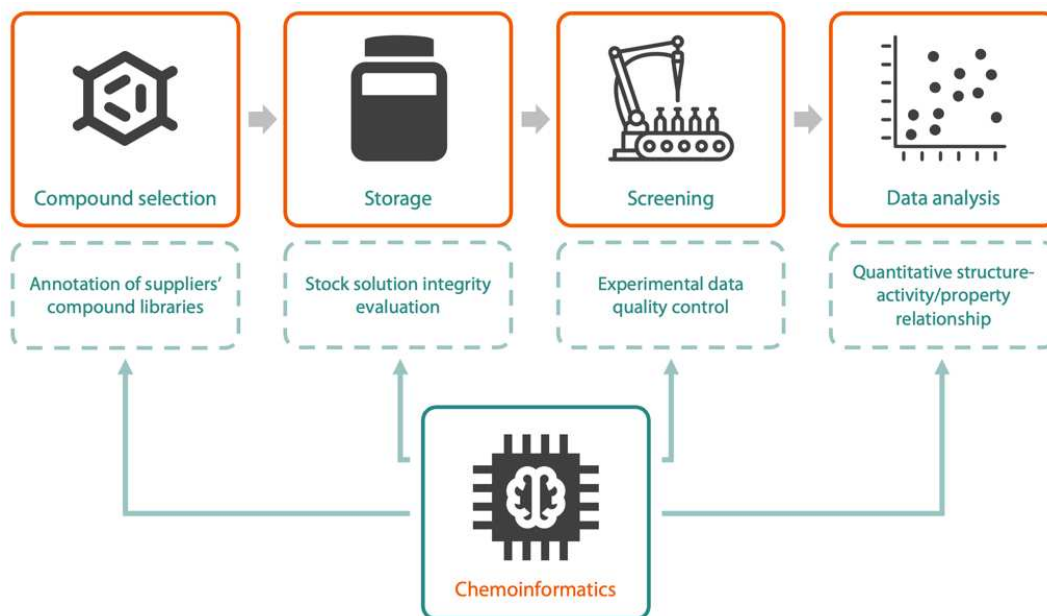


Figure 9. Overview of screening workflow and application of chemoinformatics methods at each step of the workflow.

2.3 Goal of the thesis

The high expenses associated with screening campaigns, including the use of significant time, material, and human resources, emphasize the need for cost-effective approaches. In this context, the application of chemoinformatics tools emerges as a promising solution to reduce costs and improve screening outcomes. Studies have demonstrated that incorporating chemoinformatics methods into the screening process can yield substantial benefits, with reported enhancements in hit rates ranging from 4- to 15-fold compared to random screening.^{1,50,51} These findings highlight the significant potential of chemoinformatics in maximizing the efficiency and success of screening endeavors. Consequently, my thesis is motivated by the compelling objective of developing and implementing chemoinformatics tools to support screening campaigns in the stages of data collection and analysis, quality control of screening data, development of specifically adapted predictive models, and annotation of screening libraries. The integration of these tools aims to streamline and refine the screening pipeline, addressing the challenges associated with resource consumption and maximizing the identification of potential hits.

The results presented in this thesis are organized into 3 chapters:

- Solubility
 - The solubility properties are DMSO solubility and aqueous solubility. DMSO is a standard solvent for the storage of organic compounds, and bioassays are usually performed in buffer solutions.
- Skin-related safety properties
 - The focus has been made on skin permeability and skin sensitization. Assessing skin sensitization is now a regulatory requirement for the EU under REACH Annex VII. Skin permeability is a crucial, but not easily accessible, parameter for estimating the risk of a chemical.
- ACE2 selective inhibition
 - The endpoint included in the thesis is the selective binding to angiotensin-converting enzyme (ACE2) as part of the design of biological probes capable of temporarily modulating ACE2 activity in different biological tissues and organs.

The projects presented in this thesis are the result of collaborations with various research teams and institutes, including the Plateforme Intégrée de Criblage de Toulouse (PICI) and the Plateforme de Chimie Biologique Intégrative de Strasbourg (PCBIS - UAR 3286) for solubility in DMSO and kinetic solubility in aqueous media, the Institute of Organic Chemistry and Enamine Ltd. in Kyiv (Ukraine) for selective ACE2 inhibition, and the Institut National de Recherche et de Sécurité (INRS) in Nancy for skin permeability and sensitization.

As part of this thesis, chemoinformatics solutions have been developed, including publicly available quantitative structure-activity-property relationship (QSAR/QSPR) models and user-friendly tools for deploying these models *in silico*: <https://chematlas.chimie.unistra.fr/cgi-bin/predictor2.cgi>. Models have been developed using the LIBSVM⁵² software and the KNIME² software environment. All data processing workflows have been designed to be intuitive to use and not requiring software programming expertise.

Chapter 3

QSAR/QSPR modeling methodology

This chapter provides details of different aspects of QSAR/QSPR modeling methodology applied in the thesis. QSAR/QSPR modeling involves establishment of relationship between molecules and the property or activity of interest, using machine learning (ML) algorithms. The developed QSAR/QSPR models can then be used to predict activity or property of new molecules.

3.1 Molecular standardization

Structural representation of the same molecule can often differ due to preferences of a user who uploads the data or due to possible mistakes made during registration of the data point. Different representations of the same molecule may cause an issue during reading of the molecule by a machine, as it may consider them as different chemical entities.⁵³ Therefore, it affects the quality of established structure-activity(-property) relationship during modelling, and hence its predictive performance. To avoid this, all molecular structures must be standardized according to defined rules.

ChemAxon Standardizer⁵⁴ and a KNIME² workflow, developed in our lab, were used for standardization. The standardization rules include removal of all stereochemical information, removal of solvents, removal of counterions of the main molecules, removal of explicit hydrogens neutralization of charges, dearomatization and aromatization of structures. The exact procedure depends on the project and details are provided later.

3.2 Molecular descriptors

Classical ML methods require molecular graphs to be converted into a vector of molecular descriptors. Molecular descriptors used in this work are ISIDA substructural molecular fragments (SMFs).⁵⁵ SMFs are fragmental descriptors obtained from fragmenting the molecular 2D graph and counting the fragment occurrences. Fragments are enumerated systematically from the graph using basic fragmentation schemes: sequences, atom-centered fragments and triplets (Figure 10). Sequences are strings of connected atoms and/or bonds and they correspond to the shortest possible path between each pair of atoms. Atom-centered fragments start from an atom and neighboring atoms that fall into the pre-defined topological distance (sphere) are encoded into descriptor. Triplets are all the possible combinations of 3 atoms in a graph with a defined topological distance between each pair. As part of the fragmentation process, "Atom Pairs" and "Do All Ways" were used as additional fragmentation options. "Atom Pairs" focuses on counting constitutive atoms and disregards constitutional details, while "Do All Ways" explores all pathways that connect two atoms while defining the fragments.

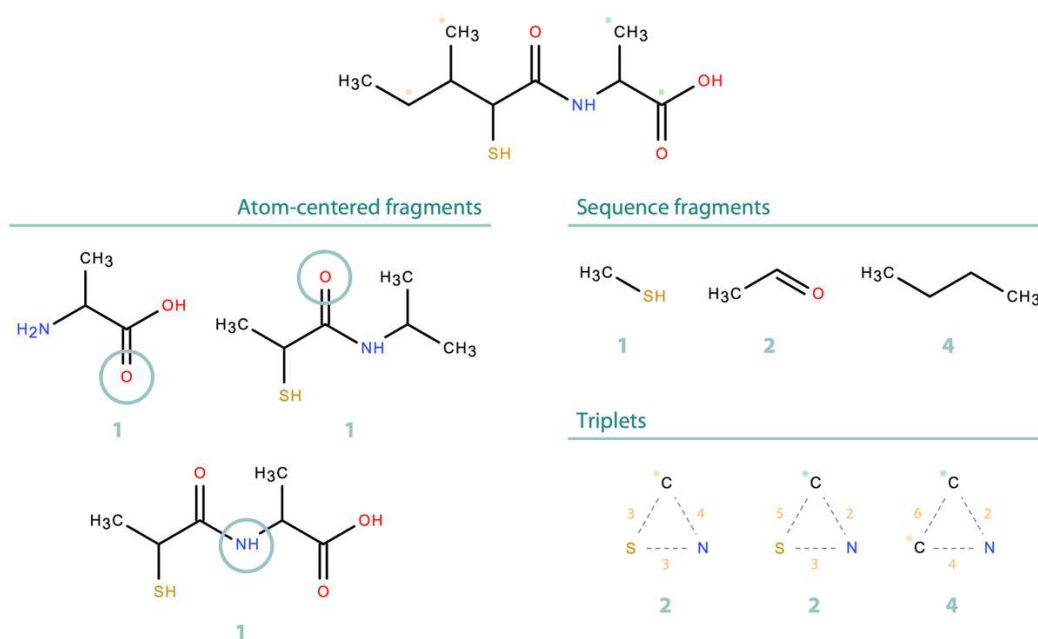


Figure 10. Example of fragmentation to ISIDA substructural molecular fragments. Stars annotated different carbon atoms. Circles highlight atom centers. The number of occurrences is given below each fragment. In triplets, the number between each atom pair indicates topological distance, or number of bonds between two atoms.

3.3 Machine learning methods

Two machine learning methods were mostly used in this work: support vector machine (SVM) for building predictive models and generative topographic mapping (GTM) for visualizing and analyzing chemical space.

3.3.1 Support Vector Machine

SVM is a supervised machine learning algorithm used for classification and regression tasks.⁵⁶ For classification tasks, the working principle is based on finding the optimal hyperplane that separates different classes of data points. The optimal position of the hyperplane is reached when the margin between the data points of different classes is maximum. Actually, the algorithm introduces soft margins that are characterized by a cost. The cost hyperparameter tunes the level of tolerance of misclassified data points. This parameter controls the balance between overfitting and underfitting. Additionally, the SVM can use the kernel formalism which is an elegant way to change the representation of the data. In this work, linear and radial basis function (RBF) kernels were used. The RBF kernel introduces a non-linearity in the modeling and the width, γ , of the RBF is an additional hyperparameter to tune. A large γ value gives more weight to each training sample. This hyper-parameter may also result into over- and under-fitting issues.

In the context of regression tasks, the SVM uses an ϵ -insensitive loss-function. The errors of the models are ignored if they are small than a user-defined ϵ threshold value. Outside of this range, they are accounted linearly, which contrasts with the vast majority of machine learning models that accounts for modeling errors in a quadratic functional.

Both classification and regression SVM models are expressing their models based on a subset of the training instances, that are termed the support vectors. These different features make the SVM attractive in the frame of this project: they are robust to small changes in the training set and to the possible presence of outliers.

In this work, SVM models were used both for prediction and outlier detection.

3.3.2 Generative Topographic Mapping

GTM⁵⁷ is an unsupervised probabilistic machine learning method used for modeling and visualization of high-dimensional data in a 2-dimensional space. This is achieved by inserting a 2D manifold into the high-dimensional descriptor space and adjusting it to align with the densest areas of the data cloud formed by molecules in the input dataset. The optimized manifold is then used to project molecules onto the 2D grid based on their closest grid nodes. The manifold is then flattened to create a 2D map. The 2D maps can be colored based on the quantitative distribution (density landscapes), the class distribution (class landscapes), and the property value distribution (property landscapes). GTM serves as a powerful tool for chemical space visualization, chemical libraries comparison, and profiling of compounds.⁵⁸⁻⁶⁰ In this work, GTM was used for visualization and analysis of the chemical space, as well as for comparison of different datasets.

3.4 Optimization algorithms

Different optimization algorithms have been employed throughout this work to find the optimal set of hyperparameters in order to maximize the objective function which is the predictive performance of machine learning models.

3.4.1 Hill climbing algorithm

Hill climbing⁶¹ is a local search optimization algorithm that aims to find the best set of hyperparameters. It starts with an initial set of hyperparameters and iteratively improves the performance by making incremental changes to maximize the predictive score of the model. At each step, it evaluates the neighboring hyperparameter values and selects the one that improves the objective function the most. The optimization loop continues until the predictive score stops improving. Hill climbing algorithm was integrated in KNIME workflows, where SVM model training for skin sensitization and skin permeability projects was performed.

3.4.2 Golden section search algorithm

Golden section search (GSS)⁶² is a local search optimization algorithm that utilizes the golden ratio to find the best set of hyperparameters within a specified interval. It works by iteratively narrowing down the search space using the golden ratio to determine the two points to evaluate the objective function. These points divide the interval into two subintervals such that the ratio of the smaller subinterval to the larger one is equal to the golden number, $(1 + \sqrt{5})/2$. By comparing the objective function values at these points, the algorithm updates the interval and continues the search until the desired precision is achieved, leading to the optimal set of variables. The GSS was used in training of SVM models for prediction of solubility of fragment-like compounds in DMSO. The GSS is actually a special case of the hill climbing algorithm.

3.4.3 Genetic algorithm

Genetic algorithm (GA)⁶³ is an evolutionary search optimization technique that aims to find the best set of hyperparameters by mimicking the process of natural selection. The algorithm starts with a population of hyperparameter sets and performs genetic operations like mutation and crossover to create new offspring. The offspring's performance, measured by the objective function (predictive performance of the model), determines their fitness. The algorithm iteratively selects the fittest individuals and by performing mutation and crossover operations, a new generation of chromosomes is generated and tested. This process continues until an optimal solution, i.e., the best hyperparameters, is found. GA was used to find optimal hyperparameters for SVM models⁶⁴ predicting ACE2 selective inhibition. The technique is best suited when more than 3 hyperparameters need to be optimized.

3.5 Evaluation metrics

3.5.1 Regression models

In this work, predictive performance of regression models is represented through determination coefficient (R^2) and root mean-squared error (RMSE) statistical metrics. R^2 is a measure of the goodness of fit of the model to the data. A typical value of R^2 ranges from 0 to 1, with a higher value indicating a better fit. However, the R^2 can take on negative values when the predicted values of a model are a worse predictor than the mean value of the target property. This situation is typically observed when considering predictions of a QSAR model out of its applicability domain (the concept of applicability domain is explained in section 3.7). Since this number has no dimension, it is often used to illustrate the models of performances of very diverse models, although this maybe sometime inappropriate. RMSE is a measure of the average prediction error of a regression model. It quantifies the difference between the predicted values and the actual values in the original units of the endpoint that is being predicted. The RMSE has the same units as the property targeted by the model. Besides, it is often proportional to the loss-function of many regression techniques, as for instances partial least squares or ridge regressions. However, SVM regressors optimize another functional.

The equations of R^2 and RMSE are provided below, where y_i^{exp} , y_i^{pred} , \bar{y}_i^{exp} , n are experimental value of i -th molecule, predicted value of i -th molecule, mean experimental value of i -th molecule, and the number of data points, respectively.

$$R^2 = 1 - \frac{\sum_i (y_i^{exp} - y_i^{pred})^2}{\sum_i (y_i^{exp} - \bar{y}_i^{exp})^2} \quad RMSE = \sqrt{\frac{\sum_i^n (y_i^{exp} - y_i^{pred})^2}{n}}$$

In the above formula, results depend on the dataset on which they are computed. Therefore, the population of the instances used to compute these performances must be provided along with the computed value.

3.5.2 Classification models

Statistical metrics used in this work to assess the predictive performance of classification models are accuracy, balanced accuracy (BA), sensitivity and specificity. True positive rate (TPR, sensitivity) measures the proportion of actual positive cases correctly identified by a model, while true negative rate (TNR, specificity) measures the same for negative class. The value of TPR and TNR ranges from 0 to 1, where 1 indicates perfect retrieval of all positive and negative class objects, respectively. Accuracy is a statistical metric that measures the overall correctness of a model by calculating the fraction of correctly predicted instances out of the total number of data points. Balanced accuracy (BA) has the same aim as accuracy, which is to provide an overall measure of the correctness of the model, while considering the proportion of positive and negative instances. It is calculated as the arithmetic average of TPR and TNR. For these two

metrics, the value of 1 indicates perfect classification; the value of 0.5 is equivalent to random guess; the value of 0 implies the opposite labelling of class objects. The equations of TPR, TNR, accuracy and BA are given below, where TP , TN , P , N are true positives (the number of correctly predicted positive data), true negatives (the number of correctly predicted negative data), total number of positives, and total number of negatives, respectively.

$$TPR = \frac{TP}{P} \qquad TNR = \frac{TN}{N}$$

$$Accuracy = \frac{TP + TN}{P + N} \qquad BA = \frac{Sensitivity + Specificity}{2}$$

3.6 Validation method

Validation of predictive performance of models was achieved either by applying the model to an external test set or by using k -fold cross-validation technique. K -fold cross-validation method involves dividing the dataset into k subsets (random sampling of data points), training the model on $k-1$ subsets, and then testing the model on the remaining subset. This process is repeated k times, with each subset serving as the test set exactly once. The predicted values made on the test of each fold are then aggregated and the performance is evaluated with respect to the original values. This technique provides an estimate of how well the model is likely to perform on unseen data from the same data distribution.

3.7 Applicability domain

The Applicability Domain (AD) is a critical concept in the development of QSAR models. It defines the domain within which a model is expected to provide reliable predictions. Models are developed using a training set of molecules, and their predictions are expected to be reliable only for molecules that are similar to those in the training set. Defining the AD of a model involves the calibration of a meta-model based on its own specific attributes and equations, and returning a "predictability score" of the molecule by the model, which is a measure of trust associated with the QSAR model output for that compound.⁶⁵

In this work, fragment control⁶⁶ is used as the AD assessment method. According to this rule, the model is not applied if a test molecule contains any new fragments that are not present in the training set. This ensures that the model is not used to make predictions for molecules that are too dissimilar to those in the training set and, as a result, improves the reliability of its predictions.

3.8 Consensus modeling

Consensus modeling is a technique that combines the predictions generated by multiple models to achieve a more accurate and robust result.⁶⁷ This approach is particularly useful when dealing with complex relationships between chemical structures and their properties that a single QSAR model may fail to accurately reflect. By aggregating the outputs of diverse models, the limitations and biases of any individual model can be mitigated, leading to improved overall performance and generalization.

In classification tasks, the consensus outcome is determined by taking the majority of votes made by each individual model, whereas for regression tasks, the consensus outcome is determined by calculating the average of predictions generated by each individual model. The individual models are trained using different pools of fragment descriptors and their predictive performance is assessed using cross-validation technique. Only models that have a k -fold cross-validation performance (BA for classification and R^2 for regression tasks) larger than a user-defined threshold are selected. The consensus models developed in this work are integrated into ISIDA Predictor software / web service.

3.9 Outlier detection

In the field of QSAR modeling, outlier detection is a crucial step in ensuring the reliability and accuracy of the models. Here, an outlier is defined as a data point that falls outside the expected range of the sample distribution.⁶⁸ Outliers can arise due to a variety of reasons, such as measurement errors, experimental variability, or the presence of compounds that exhibit unique properties not captured by the model.

The outlier identification method used in this work is based on the ensemble modeling approach.⁴⁹ The ensemble modeling approach involves applying multiple models to the fitted data and analyzing the molecules that are mis-predicted by all the models. This helps identify compounds that are anomalous and require further investigation in a unique or reduced number of modeling steps.

3.10 ISIDA Predictor software

ISIDA Predictor software is used to apply a developed QSAR/QSPR consensus model and assess the confidence of the predicted value. Prediction confidence label ("Low," "Average," "Good," or "Optimal") is based on the number of applied individual models and the consistency of their predicted values. An individual model is applied if a test molecule falls into the AD of the model (see "Applicability domain" section). Once predictions from the applied individual models are collected, a consensus prediction is

generated: the major predicted class for classification task and the average value for regression task. The output includes the predictions, prediction confidence labels, and the number of models applied. All of the best models developed in this work are included in the ISIDA Predictor⁶⁹.

3.11 Modelling workflow

The general modelling workflow applied in this thesis is given in Figure 11. The modelling details, such as the number of cross-validation folds, applied optimization algorithm, etc. are provided withing the chapter of each project.

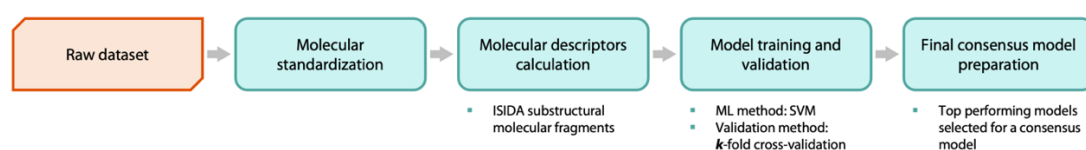


Figure 11. General modeling workflow. ML stands for machine learning.

The workflow begins with a standard data curation procedure that involves processing duplicate data points and standardizing molecular structures. Next, numerous molecular descriptor sets (ISIDA fragments) of varying topologies and lengths are generated. SVM models are trained for each descriptor set, and validation is done using the k -fold cross-validation method. Descriptor sets that show high cross-validation performance are selected, and models are fitted to the entire dataset before being included in a consensus model. All consensus models developed in this thesis are integrated into the ISIDA Predictor web service⁶⁹ and are publicly available.

3.12 KNIME workflows

Some of the models developed in this work, were generated using the workflows created using the KNIME Analytics Platform². The KNIME workflows cover the whole modelling pipeline, namely, molecular standardization, preparation of datasets for k -fold cross-validation, molecular descriptor calculation using ISIDA Fragmentor, SVM model training and validation, consensus model preparation and integration into ISIDA Predictor, model application (Table 6). The usage of the KNIME workflows require no coding knowledge and therefore are user-friendly and easy to comprehend. They were developed both for Linux and Windows operating systems. The KNIME workflows are available upon request to the Laboratory of Chemoinformatics (https://infochim.chimie.unistra.fr/?page_id=11). The figures of the workflows are provided in Appendix.

Table 6. A list of developed KNIME workflows.

Name	Description
1_standardization	Standardizes molecular structures
2_ExtCV_data_partitioning	Partitions the dataset into a defined number of sets of training and test sets for external cross-validation
3_ExtCV_descriptor_calculation	Calculates ISIDA fragment descriptors
4_ExtCV_modeling_CLS (or _REG)	Trains classification (or regression) models for external cross-validation
5_ExtCV_consensus_preparation_CLS (or _REG)	Prepares a classification (or a regression) consensus model for each fold of cross-validation
6_ExtCV_application	Applies a consensus model for the respective fold's test set
7_ExtCV_evaluation	Evaluates the predictive performance of the external cross-validation
8_final_consensus_preparation_CLS (or _REG)	Performs all the modeling steps to prepare final consensus model

Chapter 4

Solubility

4.1 Introduction

In this chapter, we discuss the significance of solubility, focusing on two critical aspects: the solubility of fragment-like compounds in DMSO and aqueous solubility. Both DMSO stock and aqueous solubilities address the same challenge - ensuring that concentrations are accurately evaluated for biological assays. Any errors related to stock or water concentration can result in assay measurement errors and ultimately lead to assay failure. Therefore, precise measurement and understanding of solubility in different conditions are essential for evaluating a drug's effectiveness and safety in drug discovery and development.

4.2 Solubility of fragment-like compounds in DMSO


4.2.1 Introduction

Ensuring the integrity and stability of DMSO stock solutions is crucial prior to proceeding to bioactivity screening. Research has demonstrated that approximately 10-20% of compounds in chemical libraries exhibit DMSO solubility below the nominal concentration.^{70,71} This can lead to a problematic "masking" effect, impairing accurate assessment of a compound's activity. The reduced concentration may arise from chemical degradation, triggered by the compound's interaction with moisture absorbed from the air.⁷² Notably, water absorption occurs during the cooling process, elevating the water content by up to 10% w/w.⁷¹⁻⁷⁴ Additionally, the solubility of compounds can be impacted by repeated "freeze/thaw" cycles when returning them to the refrigerator.⁷⁵

This study presents the development of a QSPR model specifically designed to predict solubility of fragment-like compounds in DMSO. The model can be used to effectively identify potentially insoluble molecules and minimize their occurrence within the screening library. The focus of this research centers on fragment-based screening campaigns. The findings are detailed in a published article⁴⁸, and the classification model developed during this study has been made publicly accessible through the Laboratory of Chemoinformatics' Predictor web service⁶⁹ (“Solubility in DMSO (FBS) - Classification” model in the “PhysProp” section).

Article

DMSO Solubility Assessment for Fragment-Based Screening

Shamkhal Baybekov ¹ , Gilles Marcou ¹, Pascal Ramos ², Olivier Saurel ², Jean-Luc Galzi ^{3,4} and Alexandre Varnek ^{1,*}

¹ Laboratoire de Chémoïnformatique UMR 7140 CNRS, Institut Le Bel, University of Strasbourg, 4 Rue Blaise Pascal, 67081 Strasbourg, France; sbaybekov@unistra.fr (S.B.); g.marcou@unistra.fr (G.M.)

² Institut de Pharmacologie et de Biologie Structurale, Université de Toulouse CNRS, UPS, 205 Route de Narbonne, 31077 Toulouse, France; pascal.ramos@ipbs.fr (P.R.); olivier.saurel@ipbs.fr (O.S.)

³ Biotechnologie et Signalisation Cellulaire UMR 7242 CNRS, École Supérieure de Biotechnologie de Strasbourg, University of Strasbourg, 300 Boulevard Sébastien Brant, 67412 Illkirch, France; galzi@unistra.fr

⁴ ChemBioFrance—Chimiothèque Nationale UAR3035, 8 Rue de L'école Normale, CEDEX 05, 34296 Montpellier, France

* Correspondence: varnek@unistra.fr

Abstract: In this paper, we report comprehensive experimental and chemoinformatics analyses of the solubility of small organic molecules (“fragments”) in dimethyl sulfoxide (DMSO) in the context of their ability to be tested in screening experiments. Here, DMSO solubility of 939 fragments has been measured experimentally using an NMR technique. A Support Vector Classification model was built on the obtained data using the ISIDA fragment descriptors. The analysis revealed 34 outliers: experimental issues were retrospectively identified for 28 of them. The updated model performs well in 5-fold cross-validation (balanced accuracy = 0.78). The datasets are available on the Zenodo platform (DOI:10.5281/zenodo.4767511) and the model is available on the website of the Laboratory of Chemoinformatics.

Keywords: DMSO solubility; QSPR; fragment-based screening; outlier detection; NMR



Citation: Baybekov, S.; Marcou, G.; Ramos, P.; Saurel, O.; Galzi, J.-L.; Varnek, A. DMSO Solubility Assessment for Fragment-Based Screening. *Molecules* **2021**, *26*, 3950. <https://doi.org/10.3390/molecules26133950>

Academic Editor: Martin Vogt

Received: 4 June 2021

Accepted: 23 June 2021

Published: 28 June 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Screening methods have become indisputably an integral part of the drug discovery process [1,2], from hit identification to the evaluation of pharmacological properties. Over the past decades fragment-based screening (FBS) has gained a broad acceptance as an efficient alternative to the conventional high-throughput screening (HTS) [2,3]. This is related to the core idea of FBS, which involves analysis of relatively small libraries containing simple yet diverse organic scaffolds, or fragments, and the identification of hit fragments, that will be developed into more potent lead compounds. Among the basic requirements for fragment-like compounds, well covered by the “rule of three” guidelines [4,5], solubility issues require serious attention [6,7].

Low solubility directly affects the availability of a compound in solution, which may potentially lead to masking of its actual activity. This is notably important for compounds in FBS libraries since the typical concentration of samples is around 1 mM [8–10]. Such a relatively high concentration is related to the low binding affinity of fragments, usually found in the range of μM –mM [11]. The assessment of weak ligand–target interactions, requires highly sensitive techniques such as NMR spectroscopy, etc. One of the solvents commonly used in screening methods is dimethyl sulfoxide (DMSO), a well-established standard [12].

Due to the significance of this physicochemical property, the topic of solubility prediction has been and still remains relevant. The challenge of this subject is related to the complexity of the dissolution phenomenon, which is dictated by structural features, solid state, and other physicochemical properties [13]. Very few statistical models designed to predict DMSO solubility have been reported in the literature [12,14], with only one

being publicly available [15]. Thus, Tetko et al. [15] reported a consensus model combining random forest, decision tree and Associative Neural Network individual models, trained on a large and structurally diverse dataset. However, the threshold used for categorizing compounds into “soluble” or “insoluble” classes was set to 10 mM, which is a common concentration of stock solutions.

As illustrated in Figure 1, compounds having a solubility in the range 1–10 mM, are considered soluble according to the FBS definition, but insoluble according to the stock solution definition. This means that the application of the “stock solutions” model by Tetko et al. [15] may lead to discarding compounds predicted as insoluble, but potentially suitable for FBS.

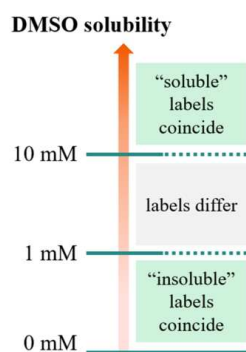


Figure 1. Solubility domains defined by the thresholds defined for stock solutions (10 mM) and FBS (1 mM). For these two threshold definitions, the “soluble”/“insoluble” labels coincide for solubility values larger than 10 mM and smaller than 1 mM, respectively. However, in the range 1–10 mM, molecules are considered soluble according to the FBS definition, but insoluble according to the stock solution definition.

This motivated us to develop a classification model predicting fragment solubility in DMSO with a categorical threshold of 1 mM. The model was built on the experimental data provided by the “Plateforme Intégrée de Criblage de Toulouse” (PICT) screening platform. During the training stage, a set of erroneous measurements were identified and removed from the PICT set. The clean dataset was then used for building SVM models. With the help of a Generative Topographic Mapping (GTM) method, the PICT dataset was compared with fragment-like compounds from the Enamine database used for the preparation of the Tetko et al. [15] “stock solutions” model. This analysis revealed some structural motifs present uniquely in PICT. The datasets collected in this work are publicly available on the Zenodo platform (DOI:10.5281/zenodo.4767511). The consensus model is freely accessible on the website of the Laboratory of Chemoinformatics (<http://infochim.u-strasbg.fr/cgi-bin/predictor2.cgi> (accessed on 16 May 2021)).

2. Data

2.1. Experimental Protocol

In order to design a fragment library for NMR-based FBS, the stock solutions of 939 fragments were prepared at a final concentration of 100 mM in DMSO-d₆, as described hereafter. The compounds, provided as powder, were dissolved at room temperature in DMSO-d₆ under vigorous shaking until solubilization. Solutions were kept overnight at room temperature, then stored at −20 °C for months. The former solutions were then used for the preparation of a set of diluted solutions with a targeted concentration of 1 mM in DMSO-d₆, to check by ¹H NMR for each fragment the chemical structure conformity and the solubility. Stock solutions at 100 mM were thawed and kept overnight at room temperature before dilution and running the NMR analysis. NMR experiments were performed

on a Bruker Avance III HD 600 MHz spectrometer (^1H Larmor frequency) equipped with a cryoprobe. NMR experiments were performed with a 30° flip angle ^1H pulse and 1.36 s of acquisition time (with a 20 ppm spectral width and a time domain 32 K complex of data points), and for each sample 32 scans were recorded with a repetition time delay of 5 s. NMR experiments were performed at 298 K and at atmospheric pressure. Quantification was performed with TopSpin, v. 3.5; Bruker Biospin software, by integration of the NMR peaks using the ERETIC2 [16] (Electronic Reference to access in vivo Concentrations) software based on the PULCON method [17]; an internal standard method which correlates the absolute intensities of spectra of compounds to be quantified with a reference spectrum. The reference spectrum was acquired as described above from a 1 mM isoleucine solution in DMSO- d_6 . The experimental error of solubility determination was estimated as 50 μM .

2.2. Data Description

The PICT dataset contained structures of 939 compounds with their corresponding DMSO concentration values ranging from 0 to 1000 μM . Since the expected concentration for DMSO samples was 1 mM, a threshold for making a division between soluble and insoluble categories was set to 1000 μM . Therefore, if concentration values were equal to 1000 μM it would be classified as soluble, and insoluble if the value was below the given threshold. Experimental error on the concentration was estimated at 50 μM ; therefore, it was decided to remove a segment of the dataset in the range 900–999 μM , as in this range the soluble/insoluble label is ambiguous. After the removal of data points with missing solubility values and the aforementioned “gray area” zone, the number of compounds in the training set was reduced to 822, where 686 and 136 compounds belonged to “soluble” and “insoluble” classes, respectively. The key physicochemical parameters varied across the PICT set in the following ranges: calculated $\log P$ -3.8 – $+3.94$, molecular weight 150–302 Da, the number of hydrogen bond acceptors 0–6, and the number hydrogen bond donors 0–3.

2.3. Data Curation

The chemical structures were standardized using a ChemAxon Standardizer [18]. Applied rules included the removal of solvents, ions, explicitly indicated hydrogen atoms, neutralization, and aromatization. All stereo labels were skipped. A detailed description of the standardization protocol is provided in Supporting Information (“Standardization protocol” section). Erroneous measurements were then detected with the help of the outlier identification procedure (see below).

2.4. Filtered Enamine Data

A subset of the fragment-like compounds was extracted from the Enamine dataset used for training of the Tetko et al. model [15] with the help of a filter, matching the same ranges of variation as the PICT dataset for ClogP, molecular weight, number of H-donors and H-acceptors. The filtering resulted in the selection of 8314 fragment-like compounds out of the initial set of 50,620 compounds.

3. Method

3.1. Molecular Descriptors

ISIDA substructural molecular fragments (SMF) [19] were used in this study. SMF descriptors are derived solely from hydrogen suppressed 2D chemical graphs. They represent fragments of different topologies (sequences of atoms and bonds, sequences of atoms only, atom-centered fragments, triplets) and size (see Table S1 in Supporting Information). The minimal length of fragments varied between 2 and 3, whereas the maximal length varied between 2 and 8. Encoding of a given sequence by its terminal atoms (“atom pairs”) was also considered. A fragment occurrence is a descriptor value. Variation of the descriptors topology; type of sequence (explicit atoms or atom pairs and

size) led to the generation of the pool of 182 subsets of descriptors. ISIDA descriptors were used in numerous QSAR studies [20–22].

3.2. Machine Learning Method

Classification models were built using the Support Vector Machine (SVM) machine learning (ML) algorithm. It was used for the selection of optimal descriptor sets, outlier identification and the generation of predictive models. The Libsvm 3.24 package [23] was used for the generation of linear SVM models. The Golden section search method was used in order to find the optimal cost parameter ranging from 0.01 to 1000 with a stopping criterion of 0.1. Optimization was performed to maximize 5-fold cross-validation (5-CV) balanced accuracy (BA).

3.3. Modeling Workflow

The modeling workflow consisted of three main stages: (1) detection of erroneous measurements, (2) selection of relevant descriptor spaces and (3) model building and implementation (Figure 2). Detection of erroneous measurements was performed following a protocol from Ruggiu et al. [24] adapted in this study to classification tasks. This approach suggests the preparation of several individual models and the identification of the common badly predicted instances. For the curated PICT dataset, 26 various fragment descriptor spaces were generated. Each subset of descriptors was used for the modeling. Five models providing the best performance in 5-fold cross-validation were selected. At the next step, common false positives and false negatives (“outliers”) detected by all selected models at the training stage were identified and inspected by the experimental team. A vast majority of them were associated with technical problems and discarded from the dataset (see “Results and discussion” section). The resulting “clean” dataset was used in a new round of model building and validation.

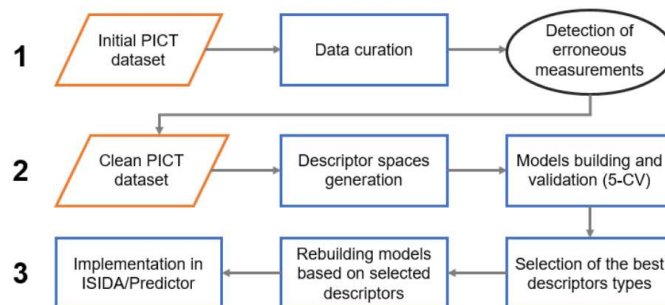


Figure 2. The modeling workflow.

At the next stage, 182 descriptor spaces were generated and used for the building and validation of SVM models. Models performing with $BA \geq 0.75$ in 5-CV were selected; the highest $BA = 0.80$ was achieved for the model based on the atom centered fragments connecting atoms pairs derived for the sequences of atoms and bonds of 3–4 atoms length (type “IIAB(3-4)_R-P”, see Table S1 in Supporting Information). Descriptors involved in the selected models were then used to develop classification models on the entire “clean” PICT dataset. Obtained in such a way, 45 individual models formed a consensus model integrated into the ISIDA Predictor tool [25]. For any new molecule, the tool assigns a solubility label according to the majority of votes for the individual models. The predictive performance of the consensus model is reasonable ($BA = 0.78$ in 5-CV). Notice that the ISIDA Predictor accounts for the fragment control [26] applicability domain (AD) of each individual model. If a new molecule is outside of the AD, the model is not applied. Along

with the predicted label, the tool provides a confidence estimation based on the ratio of the percentage models and prediction consistency.

The consensus model is freely available on the website of the Laboratory of Chemoinformatics (<http://infochim.u-strasbg.fr/cgi-bin/predictor2.cgi> (accessed on 16 May 2021)). In order to access the model, select the “PhysProp” option in the “general kind of property” section and then choose “Solubility DMSO” option from the “property to model” drop-down list. A user is invited to draw a molecule of interest or upload an SD file containing several compounds. Some screenshots illustrating the functioning of the ISIDA Predictor are given in the Supporting Information (Figure S5).

3.4. Generative Topographic Mapping

Generative Topographic Mapping (GTM) [27–30] is a dimensionality reduction method, which transforms a high-dimensional molecular descriptor space into a 2D latent space (“map”). This is achieved by introducing a 2D manifold into the high-dimensional space and adjusting a normal probability density, centered on the nodes of a rectangular grid superposed with the manifold, to the observed data distribution. Once the manifold is fitted, the compounds are projected on this 2D surface. GTM is widely used for the chemical space visualization, analysis, and compounds’ profiling [31].

Two maps were constructed: (i) for the PICT dataset and (ii) for the merged PICT and Enamine datasets. The method hyperparameters and type of fragment descriptors were optimized by maximizing the classes separation (“soluble/insoluble” for the PICT dataset and “PICT/Enamine” for the merged dataset). The compounds were encoded by atom centered fragments, including a given atom and atoms and bonds of its either 3 or 5 coordination spheres for the merged dataset and the PICT dataset, respectively. The data distribution was visualized using “class landscapes” [30], highlighting areas populated by soluble and insoluble compounds.

4. Results and Discussion

4.1. Data Visualization and Analysis

A generative topographic map built for the PICT dataset shows several clusters populated by compounds of a particular chemotype (see Figure 3). Insoluble compounds bear piperazine and morpholine fragments, soluble compounds are mostly aromatic amines, amides, piperidines and ethers, whereas compounds bearing nitro-benzene, thiophene and dihydro-thiazole fragments can be either soluble or insoluble.

A comparative analysis of the PICT and filtered Enamine datasets was performed using a generative topographic map combining both datasets. Figure 4 shows a class landscape in which the color code characterizes the presence of Enamine or PICT compounds in a particular zone of the chemical space. The map well separates blue and red zones populated by Enamine and PICT compounds, respectively, which confirms the structural diversity of the two datasets. Detailed analysis of the red zones, reveal some particular structural motifs present in the PICT and absent in the Enamine dataset (Figure 4).

4.2. Erroneous Measurements Detection

As explained above, the outliers are compounds in which the predicted labels systematically do not match the experimental ones for none of the initially developed models. There are 34 outliers which belong to three categories: experimental errors, chemical instability, and unexplained discrepancies. The list includes 31 insoluble compounds predicted as soluble and three soluble molecules predicted as insoluble (see Table S3 in Supporting Information). These modeling results were reported to the PICT team for the reassessment of experimental values. The analysis showed that 15 out of 34 potential outliers resulted from a human error during the sample preparation. Overall, during the revision of the NMR spectra, nine compounds were found to have degradation signs, whereas the values of 19 samples were likely affected by experimental errors. These 28 confirmed outliers were discarded. The remaining six compounds were claimed to have no experimental issues.

Some incorrectly predicted compounds and their correctly predicted close analogues form some sort of “solubility cliffs” (Table 1). Thus, compounds **1a** and **1b** differ by a methylene bridge between two cyclic fragments; the difference between compounds **2a** and **2b** results from the type of substituent (OH or CH₂-OH) and its position in the piperidine ring, whereas compound **3b** has two methyl groups more than the compound **3a**. These cliffs are intriguing and require further structure-activity relationship (SAR) exploration, which is beyond the scope of this work.

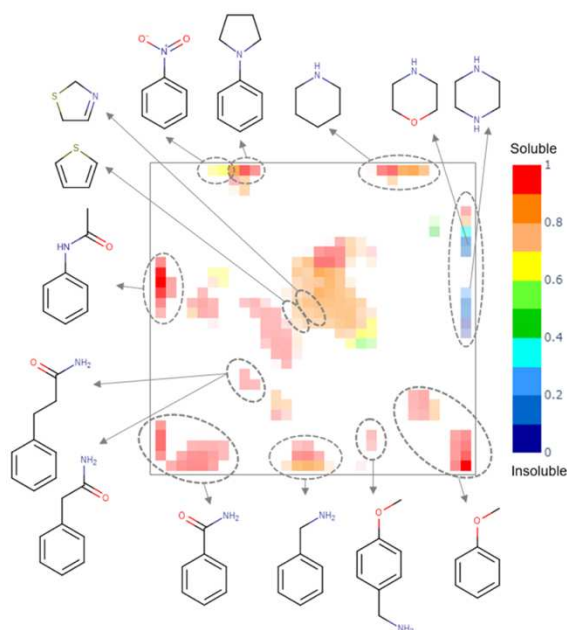


Figure 3. The class landscape for the PICT dataset. Blue and red zones are populated by insoluble and soluble molecules, respectively. Green and yellow zones contain a mixture of soluble and insoluble compounds.

Table 1. Example of incorrectly predicted compounds and their correctly predicted close analogues.

Incorrectly Predicted Compounds				Correctly Predicted Similar Compounds			
#	Compound structure	Exp	Pred	#	Compound structure	Exp	Pred
1a		Soluble	Insoluble	1b		Insoluble	Insoluble
2a		Soluble	Insoluble	2b		Insoluble	Insoluble
3a		Insoluble	Soluble	3b		Soluble	Soluble

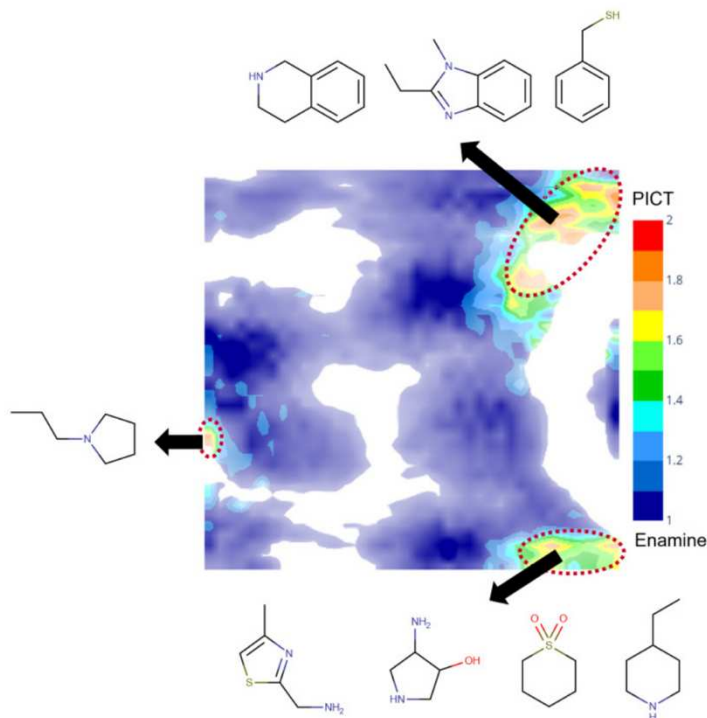


Figure 4. The class landscape depicting the coverage of a fragment-like chemical space by PICT and Enamine datasets. Blue and red zones are populated, respectively, by Enamine and PICT molecules. Green and yellow zones contain a mixture of compounds from the two datasets.

4.3. “Stock Solutions” vs. FBS Models

For the sake of comparison, the “stock solutions” model by Tetko et al. [15] was applied to the PICT dataset and, vice versa, the FBS model was applied to the Enamine dataset. Only 87.4% of the Enamine data were found inside the applicability domain of at least one FBS individual model. On the other hand, 98.6% of the PICT dataset was covered by the AD of the “stock solution” model.

Results given in Table 2 show that both models predicted soluble compounds with a high accuracy, but failed to predict insoluble ones. The latter is not surprising when the FBS model is applied to the Enamine dataset: since solubility assignment thresholds of FBS and stock solution models differ, the compounds with a solubility in the range 1–10 mM are considered soluble according to FBS and insoluble according to stock solution models. On the other hand, the compounds in which the solubility value is smaller than 1 mM are considered insoluble according to both models. This could be explained by the fact that the PICT dataset contains some unique structural motifs, e.g., thiazole, benzimidazole or tetrahydroisoquinoline (see Figure 4). It also looks like these models (at least, the “stock solution” one) are biased toward the training set composition containing mostly soluble compounds.

Table 2. Predictive performance of the FBS model on the filtered Enamine data, and of the “stock solution” model on the PICT data. The number of correctly predicted compounds with respect to the total number of compounds is given between the parentheses.

	FBS Model on Enamine Dataset	«Stock Solution» Model on PICT Dataset
Recall (soluble)	0.954 (6828/7156)	1 (676/676)
Recall (insoluble)	0.052 (6/115)	0.01 (1/101)

5. Conclusions

This work combines experimental and chemoinformatics studies of the solubility of small molecules (“fragments”) in DMSO in the context of their application in fragment-based screening. Experimentally measured data (PICT dataset) were used for the development of the first classification model for DMSO solubility fragments (FBS model). Unlike the earlier reported “stock solution” model with the categorical threshold “soluble/insoluble” of 10 mM, our model uses a more suitable threshold for fragments of 1 mM. The model displays a reasonable predictive performance in 5-fold cross-validation (BA = 0.78). Both the experimentally measured data and developed model are freely available for users.

We have demonstrated that the developed model can efficiently be used to detect erroneously measured data. Among the 28 picked compounds pointed to by the model, nine compounds were found to have degradation signs, whereas the values of 19 samples were likely affected by experimental errors.

The comparison of the PICT and Enamine datasets performed with the help of a Generative Topographic Mapping approach showed that the PICT dataset contains some unique structural motifs absent in the Enamine collection.

The results reported here demonstrate a synergism between experimental and chemoinformatics teams for obtaining, analyzing and modeling of the DMSO solubility of small molecules (“fragments”) in the context of their application in fragment-based screening.

Supplementary Materials: The following are available online: description of standardization rules, description of ISIDA fragment descriptors, description of statistical metrics, a list of models constituting the FBS consensus model, description of GTM parameters of class landscapes, a summary of predictions made on the “gray area” compounds, the outlier detection and removal workflow, a list of outliers, a list of reported classification models for the prediction of DMSO solubility and the screenshots showing the usage of the “Predictor” web-application containing our model, the PICT dataset containing experimental solubility values and class labels and the filtered Enamine dataset.

Author Contributions: Conceptualization, S.B., G.M. and J.-L.G.; methodology, S.B. and G.M.; software, S.B. and G.M.; validation, P.R. and O.S.; formal analysis, P.R. and O.S.; investigation, P.R. and O.S.; resources, P.R. and O.S.; data curation, S.B. and G.M.; writing—original draft preparation, S.B.; writing—review and editing, S.B., G.M., P.R., O.S., J.-L.G. and A.V.; supervision, G.M. and A.V.; project administration, G.M. and A.V.; funding acquisition, J.-L.G. All authors have read and agreed to the published version of the manuscript.

Funding: The fragment library and the Bruker Avance III HD 600 MHz NMR spectrometer of the Integrated Screening Platform of Toulouse (PICT) were funded by CNRS, Université Paul Sabatier, Infrastructures en Biologie Santé et Agronomie European Structural Funds, and the Midi-Pyrénées Region. This work was supported by ChemBioFrance and the Interdisciplinary Thematic Institute ITI-CSC via the IdEx Unistra (ANR-10-IDEX-0002) within the program Investissement d’Avenir. S.B. thanks the CSC Graduate School funded by the French National Research Agency (CSC-IGS ANR-17-EURE-0016).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available on the Zenodo platform (DOI:10.5281/zenodo.4767511) and in Supplementary Materials.

Conflicts of Interest: The authors declare no conflict of interest.

Sample Availability: Samples of the compounds are not available.

References

1. Proudfoot, J.R. High-Throughput Screening and Drug Discovery. In *The Practice of Medicinal Chemistry*; Wermuth, C.G., Ed.; Elsevier: New York, NY, USA, 2008; pp. 144–158, ISBN 978-0-12-374194-3. [CrossRef]
2. Farmer, B.T.; Reitz, A.B. Fragment-Based Drug Discovery. In *The Practice of Medicinal Chemistry*; Wermuth, C.G., Ed.; Elsevier: New York, NY, USA, 2008; pp. 228–243, ISBN 978-0-12-374194-3. [CrossRef]
3. Kirsch, P.; Hartman, A.M.; Hirsch, A.K.H.; Empting, M. Concepts and Core Principles of Fragment-Based Drug Design. *Molecules* **2019**, *24*, 4309. [CrossRef]
4. Congreve, M.; Carr, R.; Murray, C.; Jhoti, H. A ‘Rule of Three’ for fragment-based lead discovery? *Drug Discov. Today* **2003**, *8*, 876–877. [CrossRef]
5. Jhoti, H.; Williams, G.; Rees, D.C.; Murray, C.W. The “rule of three” for fragment-based drug discovery: Where are we now? *Nat. Rev. Drug Discov.* **2013**, *12*, 644. [CrossRef]
6. Siegal, G.; Eiso, A.B.; Schultz, J. Integration of fragment screening and library design. *Drug Discov. Today* **2007**, *12*, 1032–1039. [CrossRef]
7. Lepre, C.A. Library design for NMR-based screening. *Drug Discov. Today* **2001**, *6*, 133–140. [CrossRef]
8. Leach, A.R.; Hann, M.M.; Burrows, J.N.; Griffen, E.J. Fragment screening: An introduction. *Mol. Biosyst.* **2006**, *2*, 429. [CrossRef] [PubMed]
9. Barker, J.; Hesterkamp, T.; Whittaker, M. Integrating HTS and fragment-based drug discovery. *Drug Discov. World* **2008**, *9*, 69–75.
10. Lau, W.F.; Withka, J.M.; Hepworth, D.; Magee, T.V.; Du, Y.J.; Bakken, G.A.; Miller, M.D.; Hendsch, Z.S.; Thanabal, V.; Kolodziej, S.A.; et al. Design of a multi-purpose fragment screening library using molecular complexity and orthogonal diversity metrics. *J. Comput. Aided. Mol. Des.* **2011**, *25*, 621–636. [CrossRef] [PubMed]
11. Murray, C.W.; Rees, D.C. The rise of fragment-based drug discovery. *Nat. Chem.* **2009**, *1*, 187–192. [CrossRef]
12. Balakin, K.V.; Ivanenkov, Y.A.; Skorenko, A.V.; Nikolsky, Y.V.; Savchuk, N.P.; Ivashchenko, A.A. In Silico Estimation of DMSO Solubility of Organic Compounds for Bioscreening. *J. Biomol. Screen.* **2004**, *9*, 22–31. [CrossRef]
13. Alsenz, J.; Kansy, M. High throughput solubility measurement in drug discovery and development. *Adv. Drug Deliv. Rev.* **2007**, *59*, 546–567. [CrossRef] [PubMed]
14. Balakin, K.; Savchuk, N.; Tetko, I. In Silico Approaches to Prediction of Aqueous and DMSO Solubility of Drug-Like Compounds: Trends, Problems and Solutions. *Curr. Med. Chem.* **2006**, *13*, 223–241. [CrossRef] [PubMed]
15. Tetko, I.V.; Novotarskyi, S.; Sushko, I.; Ivanov, V.; Petrenko, A.E.; Dieden, R.; Lebon, F.; Mathieu, B. Development of Dimethyl Sulfoxide Solubility Models Using 163 000 Molecules: Using a Domain Applicability Metric to Select More Reliable Predictions. *J. Chem. Inf. Model.* **2013**, *53*, 1990–2000. [CrossRef] [PubMed]
16. Bharti, S.K.; Roy, R. Quantitative ¹H NMR spectroscopy. *Trends Anal. Chem.* **2012**, *35*, 5–26. [CrossRef]
17. Wider, G.; Dreier, L. Measuring protein concentrations by NMR spectroscopy. *J. Am. Chem. Soc.* **2006**, *128*, 2571–2576. [CrossRef] [PubMed]
18. ChemAxon Standardizer. Available online: <http://www.chemaxon.com/> (accessed on 16 May 2021).
19. Varnek, A.; Fourches, D.; Hoonakker, F.; Solov'ev, V.P. Substructural fragments: An universal language to encode reactions, molecular and supramolecular structures. *J. Comput. Aided Mol. Des.* **2005**, *19*, 693–703. [CrossRef]
20. Ruggiu, F.; Solov'Ev, V.; Marcou, G.; Horvath, D.; Graton, J.; Le Questel, J.Y.; Varnek, A. Individual hydrogen-bond strength QSPR modelling with ISIDA local descriptors: A step towards polyfunctional molecules. *Mol. Inform.* **2014**, *33*, 477–487. [CrossRef]
21. Glavatskikh, M.; Madzhidov, T.; Solov'ev, V.; Marcou, G.; Horvath, D.; Graton, J.; Le Questel, J.Y.; Varnek, A. Predictive Models for Halogen-bond Basicity of Binding Sites of Polyfunctional Molecules. *Mol. Inform.* **2016**, *35*, 70–80. [CrossRef]
22. Varnek, A.; Fourches, D.; Solov'ev, V.P.; Baulin, V.E.; Turanov, A.N.; Karandashev, V.K.; Fara, D.; Katritzky, A.R. “In silico” design of new uranyl extractants based on phosphoryl-containing podands: QSPR studies, generation and screening of virtual combinatorial library, and experimental tests. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1365–1382. [CrossRef]
23. Chang, C.-C.; Lin, C.-J. [LIBSVM]: A library for support vector machines. *ACM Trans. Intell. Syst. Technol.* **2011**, *2*, 1–27. [CrossRef]
24. Ruggiu, F.; Gizzi, P.; Galzi, J.L.; Hibert, M.; Haiech, J.; Baskin, I.; Horvath, D.; Marcou, G.; Varnek, A. Quantitative structure-property relationship modeling: A valuable support in high-throughput screening quality control. *Anal. Chem.* **2014**, *86*, 2510–2520. [CrossRef]
25. HOME—Chemoinformatics Laboratory. Available online: <http://infochim.u-strasbg.fr/> (accessed on 16 May 2021).
26. Horvath, D.; Marcou, G.; Varnek, A. A unified approach to the applicability domain problem of QSAR models. *J. Cheminform.* **2010**, *2*, O6. [CrossRef]
27. Kireeva, N.; Baskin, I.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. Inform.* **2012**, *31*, 301–312. [CrossRef]
28. Bishop, C.M.; Svensen, M.; Williams, C.K.I. GTM: The Generative Topographic Mapping. *Neural Comput.* **1998**, *10*, 215–234. [CrossRef]
29. 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. [CrossRef]

30. Horvath, D.; Baskin, I.; Marcou, G.; Varnek, A. Generative Topographic Mapping of Conformational Space. *Mol. Inform.* **2017**, *36*, 1700036. [[CrossRef](#)] [[PubMed](#)]
31. 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. [[CrossRef](#)] [[PubMed](#)]

4.2.2 Summary

To summarize, this chapter presents a comprehensive investigation of the solubility of small organic molecules ("fragments") in DMSO. The novel classification model developed herein, named the FBS model, differs from previous "stock solution" models by utilizing a more appropriate threshold of 1 mM for fragment solubility instead of the categorical threshold of 10 mM. Notably, the FBS model demonstrates promising predictive performance, with a BA of 0.78 on 5-fold cross-validation. This model can be used to identify compounds that are not feasible for FBS set-up, avoiding unnecessary expenditures. Both the new experimentally measured data from the PICT dataset⁷⁶ and the developed model⁶⁹ are freely accessible for users. The new dataset and model contribute to the broader scientific community by facilitating further research and enhancing the efficiency of compound selection for screening experiments.

4.3 Aqueous solubility

4.3.1 Introduction

Aqueous solubility is among the first properties that is screened and optimized throughout the whole drug discovery and development pipeline.^{6,77} Despite the availability of a plethora of aqueous solubility data, one of the main issues often encountered during their inspection is the lack of precise description of the experimental set-ups used to gather the data. The descriptive terms that are often used to define the nature of the solubility can be roughly resumed in two levels ontologies: solubility data types and measurement assay types (Figure 12).

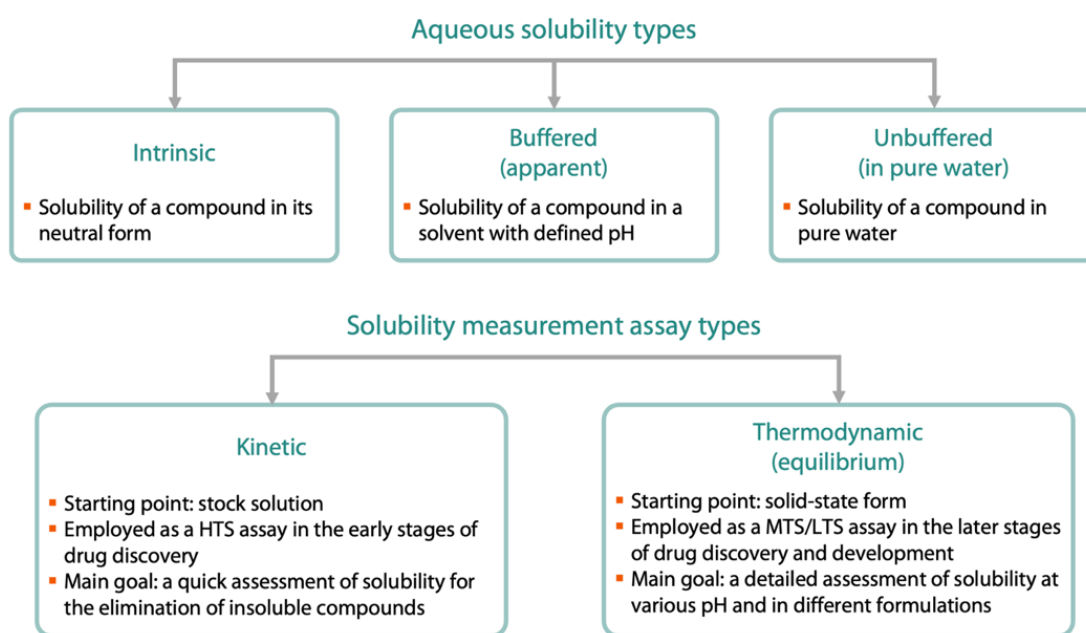


Figure 12. Types of aqueous solubility and applied solubility measurement assay. HTS, MTS and LTS stand for high-, medium- and low-throughput screening, respectively.

The first differentiation can be made based on the media where a compound is dissolved. Intrinsic solubility refers to solubility of a compound at pH when it is in its neutral form. For ionizable compounds, intrinsic solubility is measured indirectly using the CheqSol method.⁷⁸ During the measurement of buffered (or apparent) solubility the pH of a solution is defined using a buffer, whereas the pH of unbuffered solubility (or solubility in pure water) is not controlled during the experiment. For non-ionizable compounds, values of these three solubility types coincide. The buffered solubility is commonly used in screening to simulate the solubility of molecules in bio-media. The buffered solubility can be estimated from intrinsic solubility using a Henderson-Hasselbalch equation; however, this equation should be used with caution due to some limitations.^{79–81} For instance, the equation is unstable if there is no dominant micro-species at the aimed pH and depends on the number of titration sites.

The further categorization is made based on the employed measurement assays that are guided by the purpose of testing. In the early stages of drug discovery, the aim is to rapidly eliminate compounds that are not sufficiently soluble to be tested at the maximum assay concentration. Kinetic solubility is therefore favored, as it can be implemented in a high-throughput setup, involving the screening of samples prepared from stock solutions.⁸² At later stages of drug discovery and development, solubility is measured more thoroughly and tolerates a slower pace, to serve as a parameter for the bioavailability and safety of drug candidates. These solubility measurement experiments use a pure powder as a starting point and are referred to as thermodynamic solubility assays.⁸² Although both assays are important, thermodynamic solubility is more often modeled as it is considered a thermodynamic quantity, reproducible and having a direct relationship with the nature of the solute. Kinetic solubility tests, on the other hand, are less studied as they are considered to be non-reproducible, not corresponding to a thermodynamic equilibrium.

A focus has been made on differentiation of kinetic and thermodynamic solubility from a modelling point of view. The results are presented in two published articles. The kinetic solubility work is presented hereafter. The work on the challenges of accurately predicting thermodynamic solubility is published independently and was led by a colleague, Pierre Llompart, including my contributions. It is a review of the published datasets and QSAR models in the past 20 years. It emphasizes the importance of data quality and applicability domain. It also proposes a workflow of data curation of thermodynamic solubility. It has been submitted to *Scientific Data*, but still in reviewing.

The kinetic solubility paper, hereafter detailed, focuses on the repeatability and modelability of kinetic solubility assays. It explores the relationship between kinetic and thermodynamic solubility data, and examines the alignment of data from different kinetic assays. The kinetic solubility *in silico* model developed during this study was made publicly available and was uploaded to the Laboratory of Chemoinformatics' Predictor web service⁶⁹ ("Kinetic solubility - Classification" model in the "PhysProp" section). This work has been proposed to *Molecular Informatics*.

DOI: 10.1002/minf.200((full DOI will be filled in by the editorial staff))

Kinetic solubility: experimental and machine-learning modeling perspectives

Shamkhal Baybekov,^[a] Pierre Llompart,^[a,b] Gilles Marcou,^[a] Patrick Gizzi,^[d] Jean-Luc Galzi,^[c,e] Pascal Ramos,^[f] Olivier Saurel,^[f] Claire Bourban,^[d] Claire Minoletti^[b] and Alexandre Varnek^{*[a]}

Abstract: Kinetic aqueous or buffer solubility is important parameter measuring suitability of compounds for high throughput assays in early drug discovery while thermodynamic solubility is reserved for later stages of drug discovery and development. Kinetic solubility is also considered to have low inter-laboratory reproducibility because of its sensitivity to protocol parameters^[1]. Presumably, this is why little efforts have been put to build QSPR models for kinetic in comparison to thermodynamic aqueous solubility.

Here, we investigate the reproducibility and modelability of kinetic solubility assays. We first analyzed the relationship between kinetic and thermodynamic solubility data, and then examined the consistency of data from different kinetic assays. In this contribution, we report differences between

Keywords: kinetic solubility, thermodynamic solubility, comparison, QSPR

kinetic and thermodynamic solubility data that are consistent with those reported by others^[1,2] and good agreement between data from different kinetic solubility campaigns in contrast to general expectations. The latter is confirmed by achieving high performing QSPR models trained on merged kinetic solubility datasets. The poor performance of QSPR model trained on thermodynamic solubility when applied to kinetic solubility dataset reinforces the conclusion that kinetic and thermodynamic solubilities do not correlate: one cannot be used as an ersatz for the other. This encourages for building predictive models for kinetic solubility. The kinetic solubility QSPR model developed in this study is freely accessible through the Predictor web service of the Laboratory of Chemoinformatics

(<https://chematlas.chimie.unistra.fr/cgi-bin/predictor2.cgi>).

1 Introduction

Aqueous solubility is an essential property of a compound to be measured in drug discovery and development.^[3,4] It is a parameter to assess the bioavailability of a compound and it is important to avoid bias on the measurement of a bioactivity, such as a masking effect – i.e. when the saturation of an assay is due to the solubility limit of a compound and not to the biological material tested.^[5,6] Different steps of drug discovery and development focus on different aspects of solubility, which in turn dictates the choice of experimental approach used to measure solubility.^[3,4]

Solubility can be classified depending on the measurement protocol. If a setup involves the dissolution of a solid compound in a solvent, it is considered to be thermodynamic (assay) solubility. In case the source of a compound is a sample from the stock solution, the measurement is regarded as kinetic (assay) solubility. Another difference resides in the fact that thermodynamic solubility determines highest solubility limit, while kinetic determinations are carried out at a single concentration. Although kinetic solubility is operated in high throughput screening (HTS) conditions in order to anticipate solubility issues during a screening campaign, new methods have been developed during the last two decades, to also adapt thermodynamic solubility assays to HTS conditions^[1,3]. Yet, differences in experimental setups lead to several advantages of kinetic over thermodynamic measurement assays types: (i) higher dissolution rate and (ii) control of the pH. Since the starting point for kinetic solubility assays is a stock solution,

solubilization process does not involve a disruption of the crystal lattice. Nevertheless, residues of an organic solvent, which might affect the real water solubility, remain present in the final medium. The preservation of pH is ensured by the maximal concentration of the solute that is never able to compete with the buffer.

Integration of aqueous solubility data in a single dataset requires inspection of the precise definition of solubility type and measurement setup. The diversity of solubility data may be an issue if data of incompatible origins are accidentally added to a dataset for training of *in silico* models.^[7] This issue accumulates with other parameters the solubility naturally

[a] Laboratoire de Chémoïnformatique UMR 7140 CNRS, Institut Le Bel, University of Strasbourg, 4 Rue Blaise Pascal, 67081, Strasbourg, France

[b] IDD/CADD, Sanofi, Vitry-Sur-Seine, France


[c] Biotechnologie et signalisation cellulaire UMR 7242 CNRS, École supérieure de biotechnologie de Strasbourg, University of Strasbourg, 300 Boulevard Sébastien Brant, 67412, Illkirch, France

[d] Plateforme de Chimie Biologique Intégrative de Strasbourg UAR 3286 CNRS, University of Strasbourg, 300 Boulevard Sébastien Brant, 67412, Illkirch, France

[e] ChemBioFrance - Chimiothèque Nationale UAR 3035, ENSCM - 240, Avenue du Prof. E. Jeanbrau - CS 60297 - 34296 Montpellier Cedex 5

[f] Institut de Pharmacologie et de Biologie Structurale (IPBS), Université de Toulouse, CNRS, Université Toulouse III - Paul Sabatier (UT3), Toulouse, France

*e-mail: varnek@unistra.fr

 Supporting Information for this article is available on the WWW under www.molinf.com

Running title

depends on, such as solid properties (crystalline, polymorph, amorphous), particle aggregation or measurement temperature^[3], degrading the predictive performances of the models. Most of these *in silico* models are designed to predict thermodynamic solubility, whereas models predicting kinetic solubility are scarce.^[8–11] A non-exhaustive list of few reported quantitative structure-property relationship (QSPR) models targeting kinetic solubility is given in Table 1. We assume that such a small number of models is explained by a belief that kinetic solubility data are not as valuable for modelling as thermodynamic solubility data, as they are considered not reproducible due to sensitivity to experimental conditions of an assay.^[12] Nevertheless, it is kinetic solubility which is generally measured during the first stages of drug discovery and is of primary interest for screening platforms. Therefore, *in silico* models are useful upstream or in parallel to HTS and experimental kinetic solubility assessment: either for filtering compounds or to facilitate the identification and localization of problems during the assay.

Table 1. Published QSPR models predicting kinetic solubility. Performance values correspond to the highest score reported in respective articles.

Model	Availability	Performance
MetaClassifier (RF) ^[8]	No	Accuracy (test) = 0.65
Pruned MLSMR ^[9]	No	ROC AUC (test) = 0.86 MAE (test) = 0.44
GAT MTB ^[10]	No	R ² (test) = 0.3
Model10 ^[11]	No	Accuracy (test) = 0.86 ROC AUC (test) = 0.93

In this work, we investigate the reproducibility and modelability of kinetic solubility. First, we compare different kinetic methods by comparing solubility values of compounds duplicated in different datasets. Then, we analyze scatter plots comparing kinetic and thermodynamic solubility values of compounds. Finally, we report the predictive performances of models trained on kinetic solubility datasets and investigate predictions made on other kinetic solubility datasets. The best model is freely available on the web-server of the Laboratory of Chemoinformatics.^[13]

2 Data

The solubility datasets presented in this paper were used (i) to study the difference between kinetic and thermodynamic solubility assay types; (ii) to analyze the consistency between solubility data obtained using different kinetic solubility assays; (iii) to build and validate QSPR models (Table 2). Molecular structures of all the datasets were standardized using ChemAxon Standardizer^[14] (see Supplementary Information). We interpreted kinetic solubility data in terms of two classes: "Soluble" (kinetic solubility ≥ 1 mM) and "Insoluble" (kinetic solubility < 1 mM), in analogy to fragment-based screening practices^[15–17]. The precise definition of these labels needs to be adjusted depending on the specific datasets mentioned below. Table 2 resumes the

characteristics of all kinetic solubility datasets used in this work and discussed below.

2.1 Description of datasets

Among the following datasets, PICT, CNE2, Prestwick Chemicals has never been published before.

PICT

The dataset was provided by Plateforme Intégrée de Criblage de Toulouse (PICT) screening platform. It consists of kinetic solubility measurements for 939 fragments (small organic molecules). The measurements were performed in PBS buffer solution (pH 7.2) (with 1% DMSO from stock solution) using NMR technique for detection (see Supplementary Information for experimental details). Adding uncertainties in sample preparation and detection, experts recommend to interpret a fragment of this dataset as "Insoluble" if the reported concentration is < 780 μ M and "Soluble" if the concentration is > 880 μ M. In-between the solubility label is undecided. Other curation steps included removal of data points reporting a concentration greater than the nominal sample concentration (1 mM) or greater than the concentration in the stock solution, indicative of an error. After the curation and removal of 46 confirmed outliers and suspicious data points (see Supplementary Information Table S5), the total number of compounds in the dataset was 606 (513 "Soluble" and 93 "Insoluble").

Prestwick Chemicals

This dataset originates from the Prestwick Chemicals company. Kinetic solubility was measured for 1049 fragments in a buffer solution (pH 7.4) using static light scattering (SLS). Compounds are categorized as "Soluble" or "Insoluble" at 1 mM PBS (with 1% DMSO from stock solution). Data curation involved removal of identical duplicate measurements, as well as the molecules found soluble at higher concentrations, 5 mM and/or 10 mM, but not at 1 mM concentration, implying an error. The curated dataset consists of 989 compounds (900 "Soluble" and 89 "Insoluble").

Life Chemicals

Life Chemicals company provided kinetic solubility data for one of its fragment libraries^[18]. Solubility of 11457 fragments was visually determined based on scattering observed in solutions at 1 mM concentration in PBS (pH 7.4) with 0.5% DMSO. After removal of data points with no kinetic solubility, the curated dataset consists of 9276 "Soluble" molecules.

MLSMR

The Molecular Libraries Small Molecule Repository (MLSMR)^[19] is a collection of small molecules compiled under the initiative of National Institutes of Health (NIH) and screened by Sanford-Burnham Center for Chemical Genomics (SBCCG). To our knowledge, MLSMR is the largest kinetic solubility dataset available in PubChem and it is composed of 57824 data points measured in PBS (pH 7.4) using quantitative chemiluminescent nitrogen detection (CLND). Although, 0.2 mM was reported as the nominal concentration of a sample, a large fraction of the reported

Running title

concentration (about 31% of the dataset) is in the range of (0.15; 0.151]. Based on this observation, we assumed 0.15 mM as the actual sample nominal concentration and removed data points which reported concentration greater than or equal to 0.15 mM (13262 data points). Additionally, data curation included removal of duplicate molecules while taking median of their solubility values (mean of standard deviations over the duplicates = 9.85 μ M). The resulting curated dataset contained 44510 nitrogen containing compounds which are insoluble at 0.15 mM, and therefore labeled "Insoluble" at 1 mM.

Boehringer

Boehringer Ingelheim Pharma GmbH & Co. shared a dataset of 789 kinetic solubility measurements^[20] performed in PBS (pH 7.4) using nephelometry method. Data points with reported precipitate formation in DMSO stock solution and those for which solubility value was only bounded (relation denoted as ">") were removed. The curated dataset contained 605 compounds that are all "Insoluble" at 1 mM. This dataset was used for QSPR modelling. The full dataset (789 data points) was used to discuss the alignment of solubility values between different kinetic solubility assays.

CNE1 and CNE2

Chimiothèque Nationale Essentielle (CNE) is a representative collection of physical samples of pure compounds from a larger chemical library of biologically relevant substances and natural extracts called Chimiothèque Nationale^[21]. CNE1 is referring to the first generation of this representative collection of 640 compounds, most of which has been depleted. CNE2 is a currently available new representative collection of 1040 compounds. Aqueous solubility of both of these collections have been measured by the "Plateforme de Chimie Biologique Intégrative de Strasbourg" (PCBIS) screening platform. PCBIS has measured thermodynamic solubility for CNE1 collection, whereas CNE2 collection was screened for kinetic solubility. Thermodynamic solubility was measured using shake-flask method, whereas kinetic solubility was measured using HPLC-UV method, at 200 μ M nominal concentration (see Supplementary Information for details). Data curation process was identical to Oprisiu^[22]. Insoluble compounds which solubility was lower than the limit of detection have been ignored for the discussion. In addition, for CNE2, the following data points were removed:

- entries with reported concentration > 210 μ M, implying an experimental error;
- measurements with signs of impurity (multiple peaks in chromatogram);
- compounds with observed precipitation in stock solutions.

The CNE1 contains 282 compounds and the curation step yielded 525 compounds in CNE2, all of which are insoluble based on 1 mM threshold. CNE1 and CNE2 datasets were used to analyze differences between thermodynamic and kinetic solubility assay types, whereas the latter was also used for QSPR model training.

Industrial data

The kinetic solubility dataset provided by Sanofi contained solubility values of 18407 compounds measured from a 10 mM stock in PBS (pH 7.4) using nephelometry technique. The curation procedure involved duplicate molecule processing by taking median solubility value, and removal of data points in [0.8; 1.2] mM range according to expert opinion. The latter step is related to possible experimental error that could potentially change solubility label based on 1 mM threshold. The curated dataset was composed of 17320 compounds, including 71 "Soluble" and 17249 "Insoluble" compounds. A subset of the curated dataset composed of 1017 fragment-like compounds only consisted of 37 "Soluble" and 980 "Insoluble" compounds. Fragments were defined according to the rule of 3 (Ro3)^[23]: calculated logP \leq 3, molecular weight < 300 g/mol, number of hydrogen bond donors \leq 3, number of hydrogen bond acceptors \leq 3.

A subset of compounds for which both thermodynamic and kinetic solubility were measured contained 334 molecules. It was used to investigate the relationship between thermodynamic and kinetic solubility assay types. The whole dataset, "*industrial (all)*", and the fragment-like subset, "*industrial (frag)*", were used as test sets for external validation of the trained QSPR models.

2.2 Preparation of the merged kinetic solubility training set

In this section, we describe the preparation of the merged dataset comprising data of PICT, Prestwick Chemicals, Life Chemicals, Boehringer, CNE2, and MLSMR. The dataset "*industrial (all)*" and its subset "*industrial (frag)*" containing fragment-like compounds are used as external validation for QSPR models: they have been considered a posteriori, after all model building and validation has been concluded.

We identified duplicated compounds between the different datasets and tried to resolve the conflicting labels. PICT and Prestwick Chemicals have 5 compounds in common but the labels are in agreement. The labels of 2 compounds out of 27 in common between PICT and Life Chemicals datasets do not match. These 2 data were ignored because we could not resolve this conflict. There are 4 duplicates between the PICT and MLSMR datasets; labels differed for 3 of them and the discrepancy could not be solved for 1 of them - this data was ignored. For the remaining 2, the "Soluble" label was accepted because the reported concentration in MLSMR was close enough to the nominal concentration to assume that in fact, these compounds were fully dissolved.

We found 3 CNE2 molecules that had contradicting solubility class labels relative to other datasets (2 molecules between CNE2 and Prestwick Chemicals; 1 molecule between CNE2 and Life Chemicals). The 2 CNE2 molecules had solubility values (179 μ M, 180 μ M) close enough to the nominal sample concentration (200 μ M) to assume that the compounds were in fact fully dissolved, considering measurement uncertainty. For this reason, the labels "Soluble" from both Prestwick Chemicals and Life Chemicals have been accepted. The remaining CNE2 compound had "Insoluble" class label (39 μ M solubility value) which contradicted Life Chemicals' "Soluble" label. As we could not resolve this contradiction, the datapoint has not been included in the merged dataset.

Table 2. Curated solubility datasets used in this study.

Name	Compound type	Measurement technique	Max sample concentration	Curated dataset size (soluble / insoluble) ^a	Purpose			
					Kinetic vs thermodynamic solubility comparison	Kinetic solubility data reproducibility analysis	QSPR model training	QSPR model validation
PICT	Fragments	NMR	1 mM	606 (513/93)	-	+	+	-
Prestwick	Fragments	SLS	1 mM	989 (900/89)	-	+	+	-
Life Chemicals	Fragments	Visual observation	1 mM	9276 (9276/0)	-	+	+	-
MLSMR	N-containing compounds	CLND	0.15 mM ^b	44510 (0/44510)	-	+	+	-
Boehringer	Any	Nephelometry	350 µg/mL	605 (0/605)	-	+	+	-
CNE2	Any	HPLC-UV	0.2 mM	525 (0/525)	+	+	+	-
Industrial (all)	Any	Nephelometry	None	17320 (71/17249)	+	+	-	+
CNE1	Any	Shake-flask	None	282 (114/168)	+	-	-	-

^a "Soluble" and "Insoluble" labels were given according to 1 mM threshold. ^b The nominal (maximal) concentration reported in the description of the assay is 0.2 mM. NMR – nuclear magnetic resonance; SLS – static light scattering; CLND - chemiluminescent nitrogen detection; HPLC-UV – high-performance liquid chromatography-ultraviolet.

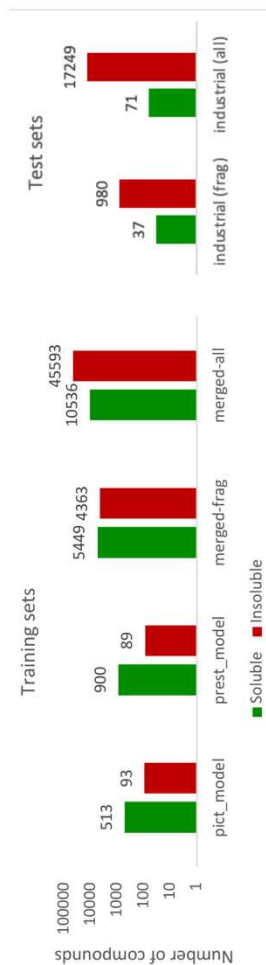


Figure 1. Distribution of "Soluble" (green) / "Insoluble" (red) classes in training and test sets. The population axis follows a logarithmic scale.

Running title

The MLSMR had 208 molecules in duplicate with the other datasets. After a thorough analysis, a large population (116 molecules) of data points was in the [140; 150] μM range, which is close enough to the nominal value of 150 μM , to assume that the compounds were in fact fully dissolved. For these 116 MLSMR data points we accepted the labels "Soluble" from the other datasets. We could not resolve the contradicting labels for the remaining 92 MLSMR duplicate measurements and these datapoints were ignored.

3 Method

3.1 Molecular descriptors

We used ISIDA substructural molecular fragments (SMF)^[24] representing 2D substructures (fragments) of various topologies (sequence of atoms only, sequence of atoms and bonds, atom-centered fragments, triplets) and sizes (see Table S1 in Supplementary Information). The descriptor value is the occurrence of a given fragment in the chemical structure. The minimal length of fragment descriptors was 2 atoms, while the maximal length varied from 2 to 5 atoms. Combination of different topologies and sizes resulted in generation of 112 descriptor sets.

3.2 Machine learning method

Support Vector Machine (SVM) method was implemented in this study for the generation of kinetic solubility QSPR models and potential outliers' detection. The SVM method offers the advantage of robustness against outliers, thanks to its epsilon-insensitive loss function. The libsvm 3.24 software package^[25] was used for training and validation of SVM models. Selection of optimal SVM hyperparameters, SVM kernels and descriptor sets was performed using genetic algorithm (GA) implemented in the libsvm-GA package^[26].

Four statistical metrics are used in our work: sensitivity, specificity, balanced accuracy (BA), Matthew's correlation coefficient (MCC). They are calculated using the equations given below (TP – true positive; TN – true negative; FP – false positive; FN – false negative). In this context, soluble class is regarded as "Positive" class, and insoluble class is regarded as "Negative" class.

$$\text{Sensitivity} = TP / (TP + FN) \quad (1)$$

$$\text{Specificity} = TN / (TN + FP) \quad (2)$$

$$\text{BA} = (\text{Sensitivity} + \text{Specificity}) / 2 \quad (3)$$

$$\text{MCC} = (TN \times TP - FN \times FP) / ((TP + FP)(TP + FN)(TN + FP)(TN + FN))^{0.5} \quad (4)$$

3.3 Modeling workflow

The modeling workflow of kinetic solubility QSPR models applied in this study can be divided into 3 steps: (1) molecular descriptor calculation; (2) model building and validation using cross-validation; (3) consensus model preparation (Figure 2). ISIDA fragment descriptors were computed for each training set during cross-validation. The hyperparameters of the models were optimized using a GA, with the cross-validation performances as scoring function. The top performing models were included in a consensus model. The selected models were then retrained on the whole dataset and included in the consensus model integrated into the ISIDA Predictor software^[27] (available both as desktop software and web

service^[13]). The ISIDA Predictor software was used to predict kinetic solubility on the industrial data. The reported external performances concern this application of the model.

In addition to the application of QSPR models, the ISIDA Predictor software incorporates an assessment of predicted value confidences. Scoring of prediction confidence is based on the number of applied models and concordance between the predicted labels given by each applied individual model of the consensus model. Each individual model prediction is considered according to the model's applicability domain, defined by fragment control rule^[28]. Fragment control states that if a test molecule contains at least one new fragment compared to those observed in the training set, the model is not applied.

The ISIDA Predictor provides 4 labels of prediction confidence: "Low", "Average", "Good", "Optimal". In this paper, for kinetic solubility QSPR models we considered only the test compounds with "Optimal" prediction confidence.

While an ideal classification model would excel at predicting compounds from both classes, in the context of kinetic solubility, the primary goal is to identify and eliminate insoluble molecules. From a statistical perspective, the model should exhibit high Specificity (the ability to predict insoluble molecules accurately) while still maintaining high BA and MCC. Performance metrics for the developed kinetic models are summarized in Table 5.

We also challenged an independent thermodynamic solubility QSPR model to predict the kinetic solubility label using a 1 mM threshold. This QSPR model has been trained on a dataset comprised of 42159 industrial and public solubility data (OCHEM, ChEMBL). The model was trained using Chemprop software package^[29] that implements a message passing neural network method. The validation performance on a test set of 5728 compounds was RMSE (root mean squared error) = 0.59.

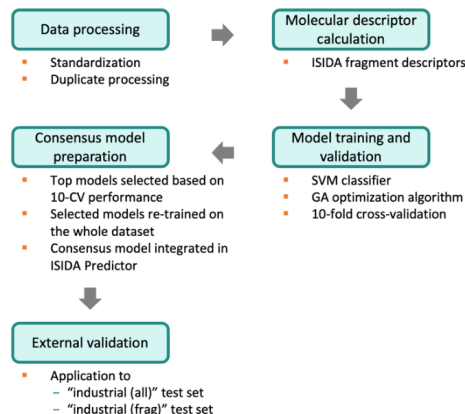


Figure 2. The modelling workflow of kinetic solubility QSPR models. The main steps are preprocessing data, computing molecular descriptor, training and validating individual models and implementation of the consensus model. External validation results from application of the final consensus model to the test sets (*industrial (all)* and *industrial (frag)* datasets). SVM – Support Vector Machine; GA – Genetic Algorithm; 10-CV – 10-fold cross-validation.

Running title

4 Results and discussion

4.1 Kinetic and thermodynamic solubility

Saal and Peterleit^[2] described three different types of relationship between kinetic and thermodynamic solubility visualized on Figure 3. The first one (Zone A) corresponds to compounds fully dissolved in a kinetic solubility measurement because their thermodynamic solubility is equal to or larger than the nominal of the measure. The second type (Zone B) is typical for the compounds whose kinetic solubility is larger than thermodynamic one. This behavior can be explained by the solid-state form of the precipitate that may differ from a kinetic to a thermodynamic measurement.^[30] In kinetic solubility measurements, the solid that forms can be amorphous or a metastable crystal polymorph; thermodynamic measurements start from a crystal that must be solubilized and are expected to let only the lowest soluble solid to form. The measurement can be complicated if the compound leads to polymorphic crystal structures.^[31] The third type (Zone C) represents compounds for whom kinetic and thermodynamic solubilities correlate.

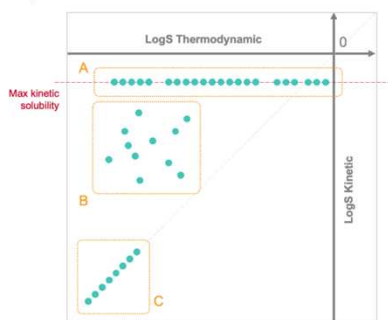


Figure 3. Different types of relationship between thermodynamic and kinetic solubility. Zone A: kinetic solubility of compounds is at the nominal concentration; zone B: kinetic solubility greater than thermodynamic solubility; zone C: kinetic solubility equals to thermodynamic solubility.

In this context two new datasets – Industrial and Chimiothèque Nationale Essentielle version 1 and 2 (CNE1 and CNE2) – have been analyzed. The former dataset shows a rather different pattern (Figure 4) from what is expected (Figure 3). The scattered data points are organized along several horizontal lines, at certain kinetic solubility values. This dataset corresponds to several kinetic solubility determination campaigns carried out at different concentrations. These horizontal lines correspond to the different nominal concentrations of the many nephelometry kinetic measurements aggregated in this dataset. The contributors to this dataset were looking for the nominal concentration at which each compound begins to appear insoluble. To this end, they scanned several of them and reported a concentration that appear to behave as in the zone A exemplified in the Figure 3.

In Figure 5, the solubility values distribution aligns with expectations (Figure 3). While the majority of data points are accumulated at about -3.7 log kinetic solubility, the others are instances of the case when kinetic solubility is greater than or equal to the thermodynamic solubility. Apart from 6 outlying data points, the overall picture resembles the pattern

described by Saal and Peterleit^[2]. The 6 compounds on the lower right hand of the plot, not matching the expectations are disclosed in the Supplementary Information (Table S4). The limit of detection at -3.7 log has been explained by Saal and Peterleit^[2] as resulting from the nominal concentration and the maximum DMSO concentration allowed in the incubation medium.

The difference between kinetic and thermodynamic solubility measurements can originate from solvent-mediated transformations occurring between different polymorphic forms of the compound.^[31,32] Recrystallization leads to the most stable polymorphic form which is characterized by its lower solubility. Measurement of a compound at any other metastable form results in different concentration (kinetic solubility) as it did not reach equilibrium state with the solution. Equally important factor is the quality of the measured compounds. Compounds with a low purity will lead to stock solutions with concentration errors, followed by calibration errors and finally, measurement errors. Additionally, it is now better understood that “kinetic solubility” does not refer to a kinetic phenomenon, and therefore, this terminology is contested.^[32]

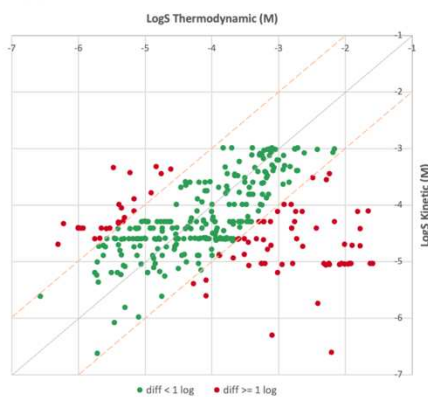


Figure 4. Comparison of kinetic and thermodynamic solubility values of the industrial dataset (334 compounds). Green dots represent differences <1 log unit between kinetic and thermodynamic values, red dots ≥ 1 log unit. Orange dashed lines show a 1 log margin.

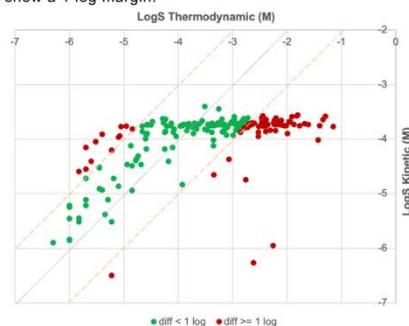
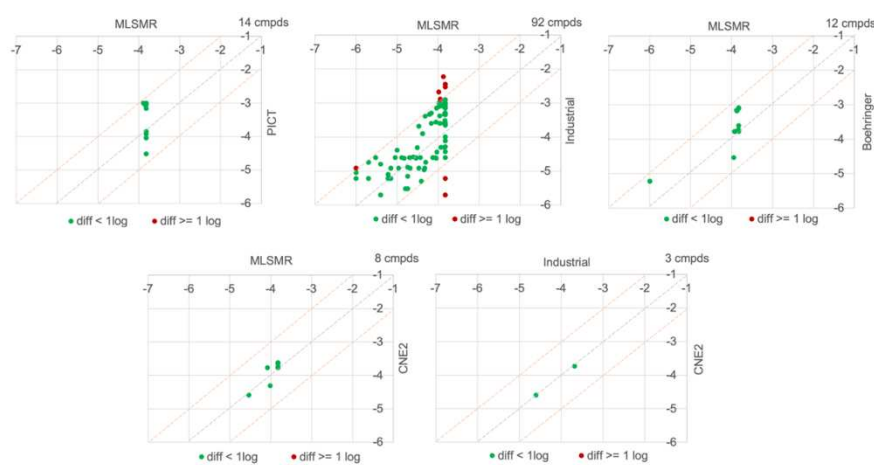


Figure 5. Comparison of kinetic and thermodynamic solubility values of Chimiothèque Nationale Essentielle dataset (186 compounds). Green dots represent differences <1 log unit between kinetic and thermodynamic values, red dots ≥ 1 log unit. Orange dashed lines show a 1 log margin.

Running title

molecular
informatics**Table 3.** The number of common compounds between each pair of kinetic solubility datasets. The LC and Prestwick datasets are composed of categorical values only, whereas the other datasets contained numerical values.

	Boehringer	LC	MLSMR	PICT	Prestwick	CNE2
LC	0					
MLSMR	12	189				
PICT	0	28	14			
Prestwick	0	39	169	5		
CNE2	0	1	8	1	5	
Industrial (all)	1	19	92	0	11	3


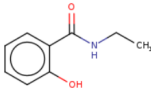
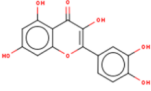
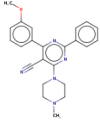
**Figure 6.** Scatter plots comparing kinetic solubility values of dataset pairs. The unit is logS (in molar). The number of common compounds is given at the top right corner of the plot. Green dots represent cases when the absolute difference between kinetic solubility values is less than 1 log unit and red dots indicate when the difference is greater than or equal to 1 log unit. Orange dashed lines show 1 log margin.

		Prestwick Chemicals		Life Chemicals		Life Chemicals	
		Soluble	Insoluble	Soluble	Insoluble	Soluble	Insoluble
PICT	Soluble	5	0	14	0	39	0
	Insoluble	0	0	2	0	0	0

Figure 7. Pairwise comparison of kinetic solubility classes for the datasets composed of fragment-like compounds.

Running title

Table 4. Comparison of kinetic solubility of compounds common to pairs of datasets. The table is composed of cases when only one compound was in common between a given pair of datasets. The case of "Industrial (all)" vs CNE2 compound is an exception: it is reported separately from the scatter plot presented in Figure 6, because it could not be quantified in CNE2 measurements.

Compound	Dataset A and solubility	Dataset B and solubility	Comment
	<u>Life Chemicals</u> Soluble at 1 mM (-3 log)	<u>CNE2</u> 0.18 mM (-3.74 log)	Difference between log values = 0.74 Good alignment (difference within 1 log unit)
	<u>PICT</u> 0.9 mM (-3.05 log)	<u>CNE2</u> 0.25 mM (-3.61 log)	Difference between log values = 0.56 Good alignment (difference within 1 log unit)
	<u>Industrial (all)</u> 0.001 mM (-6 log)	<u>CNE2</u> < 10 μM	One can assume good alignment considering the limitations of the reported data in both the industrial dataset and the CNE2.
	<u>Industrial (all)</u> 0.006 mM (-5.22 log)	<u>Boehringer</u> 0.005 mM (-5.3 log)	Difference between log values = 0.08 Good alignment (difference within 1 log unit)

4.2 Analysis of available kinetic datasets

This section reports the comparison of different kinetic solubility datasets based on common compounds between each pair of datasets. The findings of this study have been used to build the merged datasets (see section "Preparation of the merged kinetic solubility training set").

In Table 3, a number of common compounds for each pair of kinetic solubility datasets is given. The analysis of common compounds was conducted in two ways: by scatter plots, for datasets containing numerical values; by pairwise comparison of datasets containing categorical values. The cases where there was only one common compound were studied individually.

Scatter plots presented in Figure 6 generally show the good agreement between datasets (within 1 log margin). Vertical alignment of data points observed in scatter plots involving MLSMR data correspond to the upper limit of value set by the nominal sample concentration. For one compound the reported solubility is < 10 μM in CNE2 and 1 μM in the industrial dataset.

The pairwise comparison of the LC / PICT, LC / Prestwick Chemicals and PICT / Prestwick Chemicals dataset pairs shows (Figure 7) consistency of kinetic solubility data: only 2 molecules out of 16 are differently labeled in LC and PICT, in the other dataset all labels are fully consistent. The datasets whose max solubility value is less than 1 mM (MLSMR, CNE2) were not considered

during this comparison, since the solubility label between 1 mM and their nominal concentration cannot be decided.

Table 4 consists of cases where only one molecule was common to pairs of datasets, except for a compound common to "Industrial (all)" and CNE2, which was detected below the limit of quantification of the CNE2 measures. Overall, these data confirm the good agreement between kinetic solubility measures from independent sources.

4.3 Modelling of kinetic solubility

Considering the observed reproducibility of the kinetic solubility measures, we proposed to merge these datasets in order to build predictive QSPR models. For this purpose, all kinetic solubility datasets (except the industrial dataset used as an external test set) were merged in the "merged-all_model" data set. The data processing of the mixed "merged (all)" dataset resulted in 56129 molecules: 10536 "Soluble" and 45593 "Insoluble". A "merged (frag)" subset containing fragment-like compounds was prepared from the whole "merged (all)" dataset. It is composed of 5449 "Soluble" and 4363 "Insoluble" molecules, 9812 molecules in total.

Running title

molecular
informatics**Table 5.** 10-fold cross-validation (10-CV) performance of consensus QSPR models developed in this work. ^a

Model	Training set	# individual models in the consensus model	BA _{10-cv} ^b	Standard deviation (BA _{10-cv})	MCC _{10-cv} ^b	Standard deviation (MCC _{10-cv})	Sensitivity (10-CV)	Specificity (10-CV)
prest_model	Prestwick Chemicals	5	0.68	0.09	0.39	0.16	0.96	0.4
pict_model	PICT	3	0.71	0.06	0.46	0.17	0.94	0.48
merged-frag_model	Merged (frag)	7	0.87	0.01	0.75	0.02	0.91	0.84
merged-all_model	Merged (all)	12	0.93	0.004	0.86	0.005	0.88	0.98

^a Each representing ensemble of individual SVM models built on ISIDA fragment descriptors. ^b BA – balanced accuracy; MCC – Matthew's correlation coefficient.

Table 6. Performance of models on industrial kinetic solubility datasets. "industrial (frag)" is a subset of the whole "industrial (all)" dataset which is composed of only fragment-like compounds (complying Ro3).

Performance on "industrial (all)" test set					
Model	Test set size in AD after removal of common molecules (soluble/insoluble)	Sensitivity	Specificity	BA	MCC
prest_model	1004 (11/993)	1	0.73	0.87	0.17
pict_model	150 (9/141)	1	0.38	0.69	0.19
merged-frag_model	855 (19/836)	0.58	0.9	0.74	0.23
merged-all_model	345 (7/338)	0.71	0.97	0.84	0.49
therm_model	No AD filter (71/17249)	0.145	0.98	0.56	0.05
Performance on "industrial (frag)" test set					
Model	Test set size in AD after removal of common molecules (soluble/insoluble)	Sensitivity	Specificity	BA	MCC
prest_model	131 (11/120)	1	0.18	0.59	0.14
pict_model	88 (8/80)	1	0.06	0.53	0.08
merged-frag_model	195 (18/177)	0.61	0.62	0.61	0.13
merged-all_model	48 (7/41)	0.71	0.85	0.78	0.48
therm_model	No AD filter (37/980)	0.24	0.79	0.52	0.02

QSPR models built using the above datasets was compared to the models trained on individual kinetic solubility datasets. A thermodynamic solubility model has been challenged to predict the kinetic solubility classes, for comparison. Evaluation of models' performance was performed both on the whole "industrial (all)" dataset as well as its subset composed of fragment-like compounds only, "industrial (frag)". Any molecule found in both the training set and the industrial set was discarded for computing the performances: for "industrial (all)", "prest_model" training set had 8 molecules in common, "pict_model" had 0, "merged-frag_model" had 36, "merged-all_model" had 98; for "industrial (frag)", "prest_model" training set had 3 molecules in common, "pict_model" had 0, "merged-frag_model" had 36, "merged-all_model" had 37.

Since molecules in the industrial dataset are very different from the ones in the training dataset, the data coverage of all models is less than 20%: for "industrial (all)", "prest_model" was applied to 1004 molecules with "Optimal" confidence prediction label (5.8% of the "industrial (all)" with no common molecules with the training set of "prest_model"), "pict_model"

to 150 molecules (0.9%), "merged-frag_model" to 855 molecules (4.9%), "merged-all_model" to 345 molecules (2%); for "industrial (frag)", "prest_model" was applied to 88 molecules with "Optimal" confidence prediction label (12.9% of the "industrial (all)" with no common molecules with the training set of "prest_model"), "pict_model" to 131 molecules (8.7%), "merged-frag_model" to 195 molecules (19.9%), "merged-all_model" to 48 molecules (4.9%).

The results show that models trained on a combination of kinetic solubility datasets ("merged-all_model", "merged-frag_model") show higher MCC and Specificity values, compared to those trained on individual datasets, both in "industrial (all)" and "industrial (frag)" test sets (Table 6). When applied to the "industrial (frag)" test set, the "merged-frag_model" demonstrates inferior results compared to the "merged-all_model". The latter benefits from a more extensive training set, despite the former's specialization, which includes only fragment-like compounds. Moreover, one can see that the ratio of soluble to insoluble molecules in the "merged-all_model" (≈ 0.2) is closer to the ratio in the "industrial (frag)" test set (≈ 0.07), rather than the more equally distributed

9

Running title

training set of the "merged-frag_model" (≈ 1.25). Actually, the mismatch of the prior expectation of the other kinetic solubility models ("prest_model", "pict_model") compared to the actual "Soluble" / "Insoluble" distributions observed in the various dataset can have a negative impact on their performances. This adds to the weaknesses of these models resulting from the relatively small size of their training sets.

For early drug discovery solubility screening campaigns, it is better to identify and remove insoluble compounds. For this reason, it is preferable for a QSPR model to have high predictive rate of insoluble molecules (*Specificity*), while preserving a high BA and MCC. Given that, the "merged-all_model" is a better candidate to be used for virtual screening (see Supplementary Information Table S2 for details). The use of a thermodynamic solubility model for such task seems a wrong idea, as illustrated by the performance of a recent predictive QSPR model used for this task (*therm_model*, Table 6).

The benchmarking of existing models that were described in Table 1 and Table S3, is not possible due to unavailability of those models.

5 Conclusions

The analysis of kinetic and thermodynamic solubility data confirmed the previously known patterns^[2] of relationship between these two solubility types, namely, the three scenarios: (i) upper limit of kinetic solubility constrained by the assay setup, (ii) overestimation of kinetic solubility relative to thermodynamic solubility, (iii) equal kinetic and thermodynamic solubility.

Our analysis also demonstrated that the kinetic solubility data obtained using different measurement protocols are in good agreement with each other, indicating good inter-laboratory reproducibility.

This allowed us to merge the kinetic solubility data into a single dataset on which predictive models were trained. This dataset (doi:10.57745/ZWS0WC) contains exclusive data from Prestwick Chemicals, PICT and CNE2 never reported so far. The modelability of the merged dataset using different detection methods strengthen the conclusion that kinetic solubility data are not as assay-dependent as initially assumed. It should be noted that the model trained on thermodynamic solubility data fails to evaluate kinetic solubility, emphasizing that these are conceptually related but different measurements.

This contribution led to the publicly available QSPR model predicting kinetic solubility freely accessible through the Predictor web service of the Laboratory of Chemoinformatics (<https://chematlas.chimie.unistra.fr/cgi-bin/predictor2.cgi>). The model can be used for prioritization of screening compounds by preliminary assessing kinetic solubility at pH 7.4 and at 1 mM nominal concentration and a DMSO maximal concentration of 2% in the incubation medium. It is recommended to consider only "Optimal" predicted values when applying this model.

Acknowledgements

Shamkhal Baybekov thanks the CSC Graduate School funded by the French National Research Agency (CSC-IGS ANR-17-EURE-0016) for a PhD fellowship. The fragment library and

the Bruker Avance III HD 600 MHz NMR spectrometer of the Integrated Screening Platform of Toulouse (PICT) were funded by CNRS, Université Paul Sabatier, Infrastructures en Biologie Santé et Agronomie European Structural Funds, and the Midi-Pyrénées Region.

References

- [1] T. Sou, C. A. S. Bergström, *Drug Discovery Today: Technologies* **2018**, *27*, 11–19.
- [2] C. Saal, A. C. Petereit, *European Journal of Pharmaceutical Sciences* **2012**, *47*, 589–595.
- [3] J. Alsenz, M. Kansy, *Advanced Drug Delivery Reviews* **2007**, *59*, 546–567.
- [4] L. Di, P. V. Fish, T. Mano, *Drug Discovery Today* **2012**, *17*, 486–495.
- [5] L. Di, E. H. Kerns, in *Drug-Like Properties (Second Edition)* (Eds.: L. Di, E. H. Kerns), Academic Press, Boston, **2016**, pp. 61–93.
- [6] L. Di, E. H. Kerns, in *Drug-Like Properties (Second Edition)* (Eds.: L. Di, E. H. Kerns), Academic Press, Boston, **2016**, pp. 487–496.
- [7] A. Linas, I. Oprisiu, A. Avdeef, *J. Chem. Inf. Model.* **2020**, *60*, 4791–4803.
- [8] C. Kramer, B. Beck, T. Clark, *J. Chem. Inf. Model.* **2010**, *50*, 404–414.
- [9] A. L. Perryman, D. Inoyama, J. S. Patel, S. Ekins, J. S. Freundlich, *ACS Omega* **2020**, *5*, 16562–16567.
- [10] F. Broccatelli, R. Trager, M. Reutlinger, G. Karypis, M. Li, *Molecular Informatics* **2022**, *41*, 2100321.
- [11] H. Sun, P. Shah, K. Nguyen, K. R. Yu, E. Kerns, M. Kabir, Y. Wang, X. Xu, *Bioorganic & Medicinal Chemistry* **2019**, *27*, 3110–3114.
- [12] Á. Kónczöl, G. Dargó, *Drug Discovery Today: Technologies* **2018**, *27*, 3–10.
- [13] "Laboratory of Chemoinformatics - Predictor," can be found under <https://chematlas.chimie.unistra.fr/cgi-bin/predictor2.cgi>.
- [14] "Chemaxon," can be found under <https://chemaxon.com>.
- [15] A. R. Leach, M. M. Hann, J. N. Burrows, E. J. Griffen, *Mol Biosyst* **2006**, *2*, 430–446.
- [16] W. F. Lau, J. M. Withka, D. Hepworth, T. V. Magee, Y. J. Du, G. A. Bakken, M. D. Miller, Z. S. Hendsch, V. Thanabal, S. A. Kolodziej, L. Xing, Q. Hu, L. S. Narasimhan, R. Love, M. E. Charlton, S. Hughes, W. P. van Hoorn, J. E. Mills, *J Comput Aided Mol Des* **2011**, *25*, 621–636.
- [17] P. Kirsch, A. M. Hartman, A. K. H. Hirsch, M. Empting, *Molecules* **2019**, *24*, 4309.
- [18] "Fragment Library with Experimental Solubility | Fragment Libraries | Life Chemicals," can be found under <https://lifechemicals.com/fragment-libraries/soluble-fragment-library>.
- [19] "AID 1996 - Aqueous Solubility from MLSMR Stock Solutions - PubChem," can be found under <https://pubchem.ncbi.nlm.nih.gov/bioassay/1996>.
- [20] C. Kramer, T. Heinisch, T. Fligge, B. Beck, T. Clark, *ChemMedChem* **2009**, *4*, 1529–1536.
- [21] "ChemBioFrance - Chimiothèque Nationale," can be found under <https://chembiofrance.cn.cnr.fr/fr/composante/chimiotheque>.
- [22] I. Oprisiu, Modélisation QSPR de Mélanges Binaires Non-Additifs : Application Au Comportement Azéotropique, These de doctorat, Strasbourg, **2012**.
- [23] H. Jhoti, G. Williams, D. C. Rees, C. W. Murray, *Nat Rev Drug Discov* **2013**, *12*, 644–644.
- [24] F. Ruggiu, G. Marcou, A. Varnek, D. Horvath, *Mol. Inf.* **2010**, *29*, 855–868.
- [25] C.-C. Chang, C.-J. Lin, *ACM Trans. Intell. Syst. Technol.* **2011**, *2*, 1–27.
- [26] D. Horvath, J. Brown, G. Marcou, A. Varnek, *Challenges* **2014**, *5*, 450–472.
- [27] "SIDA Package – Laboratoire de Chémoinformatique – UMR 7140 CNRS," can be found under https://infochim.chimie.unistra.fr/?page_id=11.
- [28] D. Horvath, G. Marcou, A. Varnek, *J Cheminform* **2010**, *2*, O6.
- [29] K. Yang, K. Swanson, W. Jin, C. Coley, P. Eiden, H. Gao, A. Guzman-Perez, T. Hopper, B. Kelley, M. Mathea, A. Palmer, V. Settels, T. Jaakkola, K. Jensen, R. Barzilay, *J. Chem. Inf. Model.* **2019**, *59*, 3370–3388.
- [30] C. A. S. Bergström, A. Avdeef, *ADMET and DMPK* **2019**, *7*, 88–105.
- [31] H. Brittain, *American Pharmaceutical Review* **2014**, *17*, 10–16.
- [32] L. Nicoud, F. Licordari, A. S. Myerson, *Crystal Growth & Design* **2018**, *18*, 7228–7237.

4.3.2 Summary

The results obtained from the two papers shed light on the relationship between kinetic and thermodynamic solubility data and their modelability. The thermodynamic solubility paper provided insights into the failures of published models in prediction of thermodynamic solubility by emphasizing the importance of data quality, applicability domain, and careful curation of solubility datasets. The kinetic solubility paper highlights several key points. First, the kinetic solubility data obtained from different measurement protocols⁸³ demonstrate good interlaboratory reproducibility, indicating reliable agreement among the results. Second, by merging several kinetic solubility datasets, a large dataset was created⁸⁴, which was used to train a good-performing model. Finally, the fact that the QSPR model trained on thermodynamic solubility data performs poorly when applied to kinetic solubility data further confirms that although kinetic and thermodynamic solubility are conceptually linked, they represent distinct measurements.

4.4 Conclusion

This chapter provides valuable insights into the solubility properties of small organic molecules in DMSO and aqueous solutions. The newly developed FBS model for solubility in DMSO stands out by utilizing a more appropriate threshold, enhancing predictive performance, and enabling the identification of compounds unsuitable for FBS setups. Regarding aqueous solubility, the results illuminate the relationship between kinetic and thermodynamic solubility data and their modelability. The studies emphasized the significance of data quality, applicability domain, and data curation in predicting thermodynamic solubility, offering valuable guidelines for data curation and a curated AqSolDBc dataset. Furthermore, they highlighted the good interlaboratory reproducibility of kinetic solubility data and that blending of several kinetic solubility datasets into one leads to a well-performing model. The distinction between kinetic and thermodynamic solubility data underscores their unique characteristics and importance in drug discovery and development.

Both the models and the datasets used for training, for both solubility in DMSO and kinetic solubility, are freely available. These projects were achieved through a successful collaboration with Plateforme Intégrée de Criblage de Toulouse (PICT) and Plateforme de Chimie Biologique Intégrative de Strasbourg (PCBIS - UAR 3286).

Chapter 5

Skin-related safety properties

5.1 Introduction

Skin sensitization and permeability are two of the most important parameters to study in industries such as pharmaceuticals, cosmetics, and occupational safety. Understanding these parameters can help ensure better protection and handling of potentially harmful substances.

A newly developed *in vitro* skin sensitization assay, the bone marrow-derived dendritic cell (BMDC) assay, is compared with other existing *in vitro*, *in chemico*, and *in silico* tests. Additionally, a consensus classification QSAR model based on BMDC assay data is developed for preliminary assessments.

Regarding skin permeability, a new database called SkinPiX has been compiled. This database contains skin permeability coefficients and related metadata for 110 chemicals. A QSPR model has been built using the new database merged with the existing skin permeability database, HuskinDB. The newly compiled SkinPiX database, along with the developed skin permeability and skin sensitization models, can be utilized in various industries to predict and evaluate the potential skin absorption and allergenic properties of chemicals. This can lead to more informed decision-making and safer product development.

5.2 Skin sensitization

5.2.1 Introduction

Skin sensitization is a common reaction caused by repeated exposure to small molecules known as haptens. These haptens bind to skin proteins, triggering an immune response that can result in symptoms like allergic contact dermatitis. The process involves several key events, including the binding of haptens to skin proteins, the activation of skin and immune cells, that finally leads to skin sensitization.⁸ The imperative to comply with EU regulatory requirements, specifically Annex VII of REACH mandating the assessment of skin sensitization, coupled with the overarching goal of reducing animal testing advocated by REACH, propels the development of innovative *in vitro* skin sensitization methods. One such test is the bone marrow-derived dendritic cell (BMDC) assay⁹, which demonstrates potential in identifying skin sensitizing substances.

The aim of this study is to compare the predictive performance of the BMDC assay with other established *in vitro* and *in chemico* tests using common compound datasets. Furthermore, a consensus classification QSAR model based on BMDC assay data is developed for the preliminary assessment of skin sensitization. The results are published in the article provided below, whereas the model is available on the Predictor web service⁶⁹ of the Laboratory of Chemoinformatics (“Skin sensitization (BMDC) - Classification” model in the “Activity” section). The presented manuscript will be submitted to *Regulatory Toxicology and Pharmacology*. The published version may differ from this one.

Benchmarking and QSAR Modeling of BMDC Assay for Identifying Sensitizing Chemicals

Lisa Chedik*^{1#} and Shamkhal Baybekov^{2#}, Gilles Marcou², Frédéric Cosnier¹, Mélanie Mourot-Bousquenaud¹, Sandrine Jacquenet¹, Alexandre Varnek², Fabrice Battais¹

¹ Institut national de recherche et de sécurité pour la prévention des accidents du travail et des maladies professionnelles (INRS), Dept Toxicologie et Biométrie, 1 rue du Morvan, 54519 Vandoeuvre-lès-Nancy, France

² Laboratory of Chemoinformatics UMR 7140 CNRS, University of Strasbourg, Strasbourg, France

These authors contributed equally to this work. Lisa Chedik and Shamkhal Baybekov should be considered joint first author.

* Corresponding author

Abstract

The Bone-Marrow Dendritic Cell (BMDC) test is a promising assay for identifying skin and respiratory sensitizing chemicals based on the 3Rs principle.

This study expanded the BMDC benchmarking to various *in vitro*, *in chemico*, and *in silico* assays targeting different key events (KE) in the skin sensitization pathway, using common substances datasets. Additionally, a Quantitative Structure-Activity Relationship (QSAR) model was developed to predict the BMDC test outcomes for sensitizing or non-sensitizing chemicals. The modelling workflow involved ISIDA (In Silico Design and Data Analysis) molecular fragment descriptors and SVM (Support Vector Machine) classification models.

The BMDC model's performance was at least comparable to that of all ECVAM-validated models regardless of the KE considered. Compared with other tests targeting KE3 related to dendritic cell activation, BMDC was shown to have higher balanced accuracy and sensitivity with respect to both Local Lymph Node Assay and Human labels, providing additional evidence for its reliability. The consensus QSAR model exhibits promising results, correlating well with observed sensitization potential. Integrated into a publicly available web service, the BMDC-based QSAR model may serve as a cost-effective and rapid alternative to lab experiments, providing preliminary screening for sensitization potential, compound prioritization, optimization and risk assessment.

Keywords: BMDC, benchmark, LLNA, QSAR model, 3Rs principle, sensitizing chemicals

1. Introduction

Allergy to chemical substances is prevalent among both the general population and workers exposed to chemicals. Allergy can manifest through skin or respiratory symptoms, such as contact dermatitis or asthma. Allergic diseases are caused by repeated exposures to small molecules, called haptens that bind to skin or pulmonary proteins and form molecule-protein complexes, which in turn trigger an immune response. The pathophysiology of chemical allergy can be divided into two stages: a sensitization step and an elicitation step. The first phase has been extensively studied and involves different mechanisms at the molecular and cellular levels that have been described in the skin sensitization adverse outcome pathway (AOP)¹ and in a respiratory sensitization AOP which is not yet complete².

Briefly, for skin sensitization, the first step consists of the covalent binding of an hapten to skin proteins key event (KE1), followed by the induction of inflammatory response in epidermal keratinocytes (KE2), the maturation of dendritic cells (KE3) and lastly the activation, proliferation and differentiation of T-cells (KE4) which will induce the adverse effect: allergic contact dermatitis. To date, there are no validated tests for respiratory sensitization. However the development of assays (*in chemico*, *in vitro* and *in silico*) predicting skin sensitization has progressed steadily, particularly under the impetus of the ban on animal testing to identify human skin sensitizing chemicals in cosmetic products in 2009³. These different assays mostly target a specific key event in the skin sensitization AOP to allow discrimination between sensitizers and non-sensitizers. These tests have been the subject of OECD guidelines, and include but are not limited to: h-CLAT and U-SENSTM (TGD 442E), KeratinoSensTM and LuSens (TGD 442D), DPRA (TG 442C)⁴⁻⁶.

The development of robust and reliable models is a challenge given that the reference test – murine local lymph node assay (LLNA)⁷ – was an integrated test that took into account most of the steps of the AOP (KE1-KE4) from the deposition of the substance on the rodent's skin to the lymphocyte proliferation. According to the guidelines, an *in vivo* skin sensitization study, such as LLNA, should only be conducted if either the *in chemico* or *in vitro* methods cannot be utilized for the substance, or if the findings from those methods are inadequate for appropriate classification and risk assessment.

Recently, the OECD advocated a novel, defined approach (DA) for the assessment of skin sensitization that is equivalent to or more informative than the LLNA output for hazard identification. The DA involves a combination of data (*in chemico*, *in vitro* and *in silico*) that are interpreted using a mathematical model to overcome the limitations of individual tests (TGD 497)⁸. Assays proposed in this DA includes DPRA, h-CLAT, KeratinoSensTM, DEREK and QSAR Toolbox.

The BMDC (Bone Marrow-Derived Dendritic Cell) assay is one of the new *in vitro* emerging tests and appears to be promising for the identification of sensitizing substances. The BMDC model utilizes flow cytometry to measure the expression levels of phenotypic markers of mouse dendritic cells, such as MHC I, MHC II, CD40, CD54, CD80, and CD86 and therefore focuses on the KE3 of the AOP. In testing with 123 chemical compounds, the BMDC model demonstrated a high sensitivity of 94%, specificity of 78%, and accuracy of 89% compared to LLNA labelling⁹, while complying with the 3Rs principle.

Evaluating the performance of sensitization tests, such as *in vitro*, *in vivo*, *in chemico*, and *in silico*, has been challenging due to their evaluation on different datasets, making it difficult to determine the most sensitive and accurate assay and their potential value in the DA. Fortunately, the

datasets of chemicals tested in these sensitization assays often overlap and for this reason we set out to evaluate the alignment of these different assays on the same set of compounds.

In this context, our aim was firstly to assess and contrast the performance of the BMDC test with other commonly employed *in vitro*, *in chemico*, and *in silico* assays in order to investigate its advantages and limitations and to gain a comprehensive understanding of its capabilities. The tests considered in the comparison included KeratinoSens™, LuSens, h-CLAT, mMUSST, U-SENS™, DPRA, and Pred-Skin. To ensure fairness, we used datasets that contained the same compounds for all the tests. The assessment encompassed the ability to predict both human skin sensitization and LLNA labels.

To capitalize on the substantial amount of robust data generated from the BMDC test and maximize its impact for chemical safety assessment purpose, a companion QSAR (Quantitative Structure-Activity Relationship) model was developed in the second part. This model aims to predict the sensitization potential of chemical compounds based on the BMDC test results. Its design, performance, benefits and availability to the scientific community via a free web service are discussed in this article.

2. Methods

2.1 Data sources

BMDC assay data were obtained at the French National Research and Safety Institute for the Prevention of Occupational Accidents and Diseases (INRS) and were recently published⁹. These are the expression data of 6 specific markers of dendritic cells activation and the results of the sensitizing potential assessment analysis for 123 substances. The chemicals tested with the BMDC model were all purchased with a very high chemical purity (over 95%) to avoid any impurities being responsible for the observed effects. The sensitization labels for LLNA were extracted from Battais et al.'s publication⁹.

The Integrated Chemical Environment (ICE) was designed to facilitate the development, evaluation, and application of new approach methodologies. It provides a number of datasets focusing on toxicity endpoints¹, all of which were compiled and curated by National Toxicology Program Interagency Center for the Evaluation of Alternative Toxicological Methods (NICEATM)². In this study, we used the skin sensitization dataset, originally consisting of over 24,000 data entries for nearly 2,000 unique compounds. This dataset includes data from both animal and non-animal tests and served as a valuable source of data for DPRA, KeratinoSens™, LuSens, h-CLAT, U-SENS™ and mMUSST assays.

Additionally, human data were extracted from the Pred-Skin dataset. The training sets used to build models which constitute Pred-Skin software¹⁰ are available on the GitHub page of the project³. The KNIME workflow used to process the data to compare assay's performances are accessible in the online repository (<https://doi.org/10.57745/PPAMKY>) for transparency.

2.2 Benchmarking

While each skin sensitization assay had its own specific dataset, they were all applied to commonly studied chemical compounds related to sensitization, leading to overlap among the datasets used in different assays. Therefore, overlapping data sets were systematically identified and

¹ ICE: Integrated Chemical Environment: <https://ntp.niehs.nih.gov/whatwestudy/niceatm/comptox/ct-ice/ice.html> (accessed 2 November 2022)

² ICE Data Sets: <https://ice.ntp.niehs.nih.gov/DATASETDESCRIPTION> (accessed 13 October 2022)

³ Training sets of Pred-Skin models <https://github.com/joyvb/Pred-Skin/tree/master/Datasets> (accessed 3 November 2022)

statistics on these overlapping cases only were calculated. The list of compounds considered in this study and their label as sensitizers or non-sensitizers in humans or according to different tests is presented in **Supplemental 1**.

The following performance measures were reported:

$$\text{Accuracy: } (TP + TN) / (P + N) \quad (1)$$

$$\text{Balanced Accuracy (BA): } TP / (TP + FN) + TN / (TN + FP) \quad (2)$$

$$\text{Sensitivity: } TP / (TP + FN) \quad (3)$$

$$\text{Specificity: } TN / (TN + FP) \quad (4)$$

Notations: TP – true positives; TN – true negatives; FP- false positives; FN- false negatives; P – total number of positives; N – total number of negatives; positives are sensitizers and negatives are non-sensitizers.

2.3 QSAR modeling

An overview of the modelling workflow used to generate our QSAR model is shown in **Figure 1**. It starts with data curation of molecular structures, followed by training and validation of models built on various molecular descriptor sets. The final step is the preparation of a consensus model that is integrated in the ISIDA (In Silico Design and Data Analysis) Predictor web-service⁴ and is publicly available for users.



Figure 1. The overview of a Quantitative structure-activity relationship (QSAR) modelling workflow.

All the QSAR modelling steps were executed with KNIME Analytics Platform¹¹. Support Vector Machine (SVM) models were trained and applied using the LIBSVM package¹². Data points were sampled randomly (with a fixed seed) at all steps involving data partitioning. Details on model building and validation are reported in **Supplemental 2**.

ISIDA Predictor software was used to apply a developed consensus QSAR model to a set of molecules. Prediction confidence label (“Low”, “Average”, “Good”, “Optimal”) was assigned based on the number of applied models and the consistency among the predicted values of applied individual models. The individual models that constitute the consensus model were applied if a test molecule passed the fragment control applicability domain (AD) filter¹³. A test molecule was considered out of AD if it contains a substructural fragment that had not been encountered in the training set of the individual model and therefore ISIDA Predictor software did not apply this individual model. Once predictions made by the applied individual models were collected, a consensus prediction was derived by taking the value that was predicted by most of the individual models. The output consists of predictions, prediction confidence labels and number of applied models.

3. Results and Discussion

3.1 BMDC benchmarking with other sensitization assays

Evaluating diverse skin sensitization tests (*in vitro*, *in vivo*, *in chemico*, and *in silico*) is challenging due to varying datasets, hindering the identification of the most sensitive and accurate assay. A benchmark offers an opportunity for comprehensive evaluation and comparison of these

⁴The web-service of the Laboratory of Chemoinformatics: <https://chematlas.chimie.unistra.fr/cgi-bin/predictor2.cgi> (accessed 31 August 2023), “Skin sensitization (BMDC) - Classification” model in the “Activity” section.

assays, assessing their strengths and limitations in terms of sensitivity, specificity, accuracy, and reproducibility. This informs the selection of reliable tests and guides the development of integrated strategies for improved predictive accuracy and reliability.

Any sensitization test aim to predict the effects observed in human. However, LLNA is considered as the gold standard tool for predicting human sensitization as it provides a large source of quantitative data on *in vivo* sensitization. As such, it is considered as the reference method during the development and comparison of new alternative methods.

In their study, Battais et al. (2023) highlighted the strong performance of the BMDC test in comparison to LLNA labels (accuracy: 0.89, sensitivity: 0.93, specificity: 0.78), demonstrating a good alignment between BMDC and LLNA results⁹.

Through the concatenation of the BMDC, ICE, and Pred-Skin datasets (**Supplemental 1**), we successfully identified the compounds that received accurate predictions and those poorly predicted by various tests targeting different KEs in the skin sensitization AOP. This comprehensive analysis allowed for a thorough evaluation of the assay performance. Pairwise comparisons of performances of the BMDC test with other models with respect to LLNA label (DPRA n=99, h-CLAT n=91, KeratinoSens™ n=105, LuSens n=45, mMUSST n=35, Pred-Skin n=118, U-SENS™ n=97) have revealed that BMDC has a systematically better balanced accuracy (**Figure 2**).

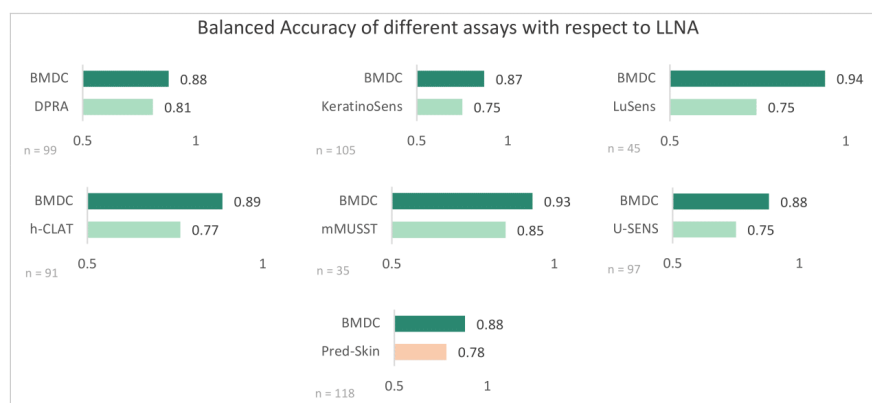


Figure 2. A pairwise comparison of Bone-Marrow Dendritic Cell (BMDC) assay with other assays with respect to Local Lymph Node Assay (LLNA) labels. Performance metric is balanced accuracy. At the bottom left part of each bar plot a number of common compounds with labels of both assays that were compared is given. Dark green bars represent BMDC performance; light green bars represent other *in vitro* and *in chemico* methods; an orange bar represents *in silico* method.

The results of DPRA, KeratinoSens™, h-CLAT, Pred-Skin, and BMDC with respect to LLNA class were analyzed together (**Table 1**). Among the commonly tested compounds (n=74), the BMDC assay exhibited better performance than other tests for all metrics (balanced accuracy, accuracy, sensitivity and specificity). The h-CLAT test, on the other hand, performed the worst in terms of BA (0.76), accuracy (0.82) and specificity (0.63), while Pred-Skin had the lowest sensitivity (0.86) among the five tests.

Table 1. Performances of skin sensitization assays with respect to Local Lymph Node Assay (LLNA) labels. The evaluation was performed on a set of common compounds (n = 74). Balanced accuracy

(BA) of Bone-Marrow Dendritic Cell (BMDC) test is high and close or similar to the other *in vitro* and *in silico* assays.

Assay	DPRA	KeratinoSens™	h-CLAT	Pred-Skin	BMDC
Data source	ICE	ICE	ICE	Pred-Skin server	Battais 2023
BA	0.84	0.81	0.76	0.80	0.89
Accuracy	0.87	0.85	0.82	0.82	0.91
Sensitivity	0.89	0.89	0.89	0.86	0.93
Specificity	0.79	0.74	0.63	0.74	0.84

In comparison to other assays targeting skin sensitization KE3 (mMUSST, h-CLAT, U-SENS™, and BMDC) on the same 30 substances, BMDC consistently showed higher BA (0.96), accuracy (0.97) and sensitivity (1) than other tests (**Supplemental 3**). The mMUSST assay had the best specificity (1) but at the cost of the worst performance in sensitivity (0.74) and accuracy (0.83). The h-CLAT and the U-SENS™ tests also displayed the worst performances in terms of balance accuracy and specificity, respectively (0.86) and (0.82), which is consistent with previously reported data in TGD442E, where a specificity of 0.66 and 0.65 was mentioned, respectively⁴. These findings suggest that BMDC is a reliable *in vitro* model for predicting the outcome of the KE3 of the AOP of skin sensitization and is a strong for inclusion among the list of validated tests for this key event, along with GARD, KeratinoSens™ and h-CLAT.

There were 12 out of 118 compounds mispredicted by BMDC assay compared to the LLNA labels (**Supplemental 1**). Among these, three compounds (with CAS number given in square brackets) (2-hydroxypropyl methacrylate [27813-02-1], propylparaben [94-13-3] and benzyl benzoate [120-51-4]) were incorrectly predicted by the majority of assays. Additionally, BMDC specifically mispredicted 8 compounds in relation to LLNA: ethylvanillin [121-32-4], coumarin [91-64-5], 6-methylcoumarin [92-48-8], 3-chloro-4 methoxybenzaldehyde [4903-09-7], diethylacetaldehyde [97-96-1], abietic acid [514-10-3], dihydro coumarin [119-84-6], squaric acid [2892-51-5].

Furthermore, among the 118 compounds tested with the BMDC assay, 54 compounds were incorrectly predicted by one or more other assays, except BMDC, with respect to LLNA label (**Supplemental 1**). Ten instances were correctly predicted by BMDC but mispredicted by at least three different assays with respect to LLNA labels: 2-acetylcyclohexanone [874-23-7], 1-bromohexane [111-25-1], 1-iodohexane [638-45-9], benzocaine [94-09-7], methyl paraben [99-76-3], methoxyacetophenone [100-06-1], benzoyl peroxide [94-36-0], hexyl cinnamic aldehyde [101-86-0], 2,2,6,6-tetramethyl-3,5-heptanedione [1118-71-4] and squaric acid diethyl ester [5231-87-8].

It is worth noting that Pred-Skin's performance with respect to LLNA classification was surprisingly low (BA 0.77), likely due to its training on human sensitization labels, which may not fully align with actual LLNA labels. Indeed, based on the ICE dataset of 82 compounds that had both LLNA and human sensitization labels (**Supplemental 1**), 9 mispredicted compounds by LLNA with respect to human labels, either false positives and false negatives, can be identified: sulfanilamide [63-74-1], benzocaine [94-09-7], coumarin [91-64-5], hexyl cinnamic aldehyde [101-86-0], pentachlorophenol [87-86-5], 2-methoxy-4-methylphenol [93-51-6], alpha-methyl cinnamaldehyde [101-39-3] and benzyl benzoate [120-51-4], cyclamen aldehyde [103-95-7]. A detailed scrutiny of 10 incorrectly predicted compounds by LLNA with respect to human data was already discussed in previously published papers^{13-15,17-24}. Despite that, the performances of LLNA assay with respect to human sensitization classification were good (BA: 0.86, accuracy: 0.89, sensitivity: 0.95, specificity: 0.77) which confirmed previous reports that LLNA and human sensitization data align relatively well^{16,19,21,22}.

In order to overcome potential biases linked to the LLNA classification and compare the performances of the various tests with respect to true human sensitization labels, pairwise comparisons of BMDC with other models' performances were conducted (DPRA n=71, h-CLAT n=73, KeratinoSens™ n=73, LuSENS n=39, mMUSST n=31, Pred-Skin n=82, U-SENS™ n=67). The results showed that BMDC consistently achieved higher or equal balanced accuracy, except for the Pred-Skin *in silico* approach, which had better performances (0.98) (Figure 3). However, it is essential to note that all compounds included in the Pred-Skin dataset (n=82) are also present in the training set of the *in silico* model, leading to an overestimation of performances in real situations. To provide a fair assessment of Pred-Skin's performances, an analysis of these predictions on an independent dataset would be necessary.

When comparing the results of several models (DPRA, KeratinoSens™, h-CLAT, Pred-Skin, BMDC) with respect to human class on a dataset of common tested substances (n=62), the BMDC assay demonstrated equal or superior performance in terms of balanced accuracy compared to other experimental tests, except for the *in silico* Pred-Skin approach (BA = 0.99, accuracy = 0.98, sensitivity = 0.98 and specificity = 1) for the reason given above (Table 2). On the other hand, the h-CLAT assay showed the poorest performance in terms of balanced accuracy (0.74), accuracy (0.81), and specificity (0.6). The lowest sensitivity values was reported for DPRA (0.85).

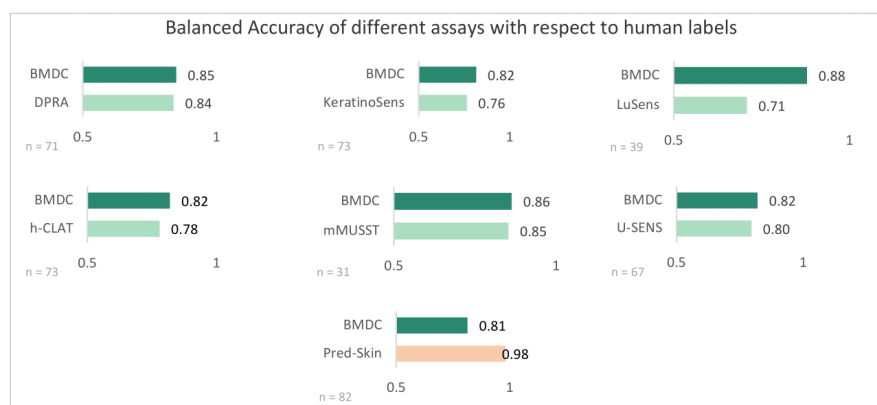


Figure 3. A pairwise comparison of Bone-Marrow Dendritic Cell (BMDC) assay with other assays with respect to human labels. Performance metric is balanced accuracy. At the bottom left part of each bar plot a number of common compounds with labels of both assays that are being compared is given.

Dark green bars represent BMDC performance; light green bars represent other *in vitro* and *in chemico* methods; an orange bar represents *in silico* method.

Table 2. Performances of skin sensitization assays with respect to human labels. The evaluation was performed on a set of common compounds (n = 62).

Assay	DPRA	KeratinoSens™	h-CLAT	Pred-Skin	BMDC
Data source	ICE	ICE	ICE	Pred-Skin server	Battais 2023
BA	0.83	0.84	0.74	0.99	0.87
Accuracy	0.84	0.86	0.81	0.98	0.90
Sensitivity	0.85	0.87	0.87	0.98	0.94
Specificity	0.8	0.8	0.6	1	0.8

Similarly, BMDC showed equal or better performances compared to other KE3 targeting assays (mMUSST, h-CLAT, U-SENS™) in predicting human class on a dataset of common tested substances (n=27) (**Supplemental 4**). Although mMUSST showed high specificity (1), it exhibited the lowest sensitivity (0.75) among the four assays, whereas the BMDC assay achieved the highest sensitivity (0.94). The U-SENS™ assay shared the same sensitivity (0.94) but at the cost of the lowest specificity (0.73). The BMDC assay emerged as one of the top-performing models for forecasting the KE3 outcome.

Among the 12 compounds mispredicted by the BMDC test compared to human labels (**Supplemental 1**), seven compounds were also misclassified by the LLNA test (hexyl cinnamic aldehyde [101-86-0], pentachlorophenol [87-86-5], 2-methoxy-4-methylphenol [93-51-6], alpha-methyl cinnamaldehyde [101-39-3], benzocaine [94-09-7] and sulfanilamide [63-74-1] and cyclamen aldehyde [103-95-7]). This is probably related to the fact that both methods are based on a mouse model which has its own specificities in terms of toxicokinetics (percutaneous absorption and metabolism) and immune response. From the remaining five compounds, three (propylparaben [94-13-3], dihydrocoumarin [119-84-6] and abiestic acid [514-10-3]) were also poorly predicted by the other *in vitro* assays. Propylparaben [94-13-3] was identified as a non-sensitizer by LLNA and various human sensitization tests, as well as several other *in vitro* (mMUSST) and *in chemico* (DPRA) assays^{23,24}. However, some *in vitro* assays (BMDC, h-CLAT, SENS-IS, U-SENS™, GARD, and KeratinoSens™) predicted it to be a weak sensitizer, and some human patch tests classified it as a very rare sensitizer²⁴. The discrepancies could be due to differences in metabolic capacity among the various *in vitro* assays. Abiestic acid is considered a sensitizer by the LLNA assay and some human data, but it is generally not classified as such and is labeled as a pre-hapten, suggesting a potential sensitization route through hydroperoxides produced during autoxidation²⁵.

Ethylvanillin [121-32-4] and 6-methylcoumarin [92-48-8] were only mispredicted by the BMDC test. For ethylvanillin, an explanation could be the generation of a Schiff base/quinone precursor alert and is classified as a pre/pro hapten²⁶.

Among the compounds correctly predicted by the BMDC test in relation to human sensitization labels, we found 28 instances that were incorrectly predicted by other assays, with methoxyacetophenone [100-06-1] and benzoyl peroxide [94-36-0] being mispredicted by the highest number of assays (**Supplemental 1**).

In absence of human labels, it is complicated to conclude on some chemicals. They are either mispredicted by the BMDC assay with respect to LLNA label (3-chloro-4-methoxybenzaldehyde [4903-09-7], diethyl acetaldehyde [97-96-1], squaric acid [2892-51-5] and 2-hydroxypropyl methacrylate [27813-02-1]) or accurately predicted by the BMDC assay in comparison to the LLNA assay, but mispredicted by at least three other sensitization tests (1-iodohexane [638-45-9], 2-acetylcyclohexanone [874-23-7], methyl paraben [99-76-3], 2,2,6,6-tetramethyl-3,5-heptanedione [1118-71-4], and squaric acid diethyl ester [5231-87-8]).

3.2 QSAR consensus model to predict the sensitization potential of chemical according to the BMDC assay.

The pre-processing procedure (as described in Supplemental 2) was conducted to standardize the molecular structures of the BMDC dataset, resulting in 118 chemicals. The BMDC skin sensitization label of 2-methyl-4-isothiazolin-3-one [2682-20-4] was found different for two chemical providers of the same compound. For this reason, it was excluded from the training set, resulting in 117 compounds

with 83 identified as sensitizers and 34 as non-sensitizers according to the BMDC assay. To develop a robust predictive model, 2300 individual models were trained using Support Vector Machine (SVM) algorithm and ISIDA molecular descriptors. The best performing models were combined into a consensus model, which outperformed the individual models and included applicability domain for prediction confidence.

The QSAR consensus model, comprised of 45 individual models, demonstrated high internal cross-validation BA values ranging from 0.94 to 1 (Table 3). During validation, the consensus model showed high predictive performance, achieving a BA of 0.82 for the highest confidence prediction (“Optimal”). Considering lower performance predictions (“Optimal” and “Good”) increases the coverage of predicted compounds from 32% to 70%, but at the cost of a lower correct prediction rate (BA from 0.82 to 0.81).

Table 3. A 5-fold external cross-validation performance of the consensus model built on Bone-Marrow Dendritic Cell (BMDC) class data. “Optimal” and “Good” are prediction confidence labels. BA stands for balanced accuracy.

Predictions	Total	Sensitizers	Non-sensitizers	BA	Accuracy	Sensitivity	Specificity
All	118	83	34	0.77	0.77	0.77	0.765
“Optimal” or “Good”	83	61	22	0.81	0.83	0.85	0.77
“Optimal”	38	26	12	0.82	0.84	0.885	0.75

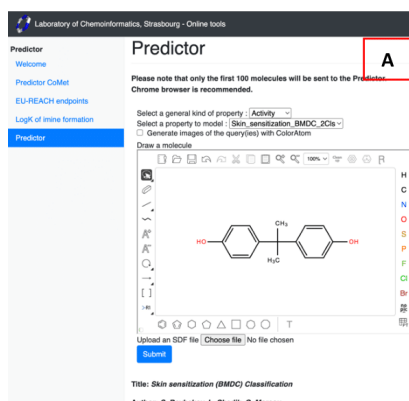
6 compounds were mispredicted while the prediction confidence was “Optimal”: 3 BMDC non-sensitizers predicted as sensitizers (vanillin [121-33-5], nonanoic acid [112-05-0], chlorobenzene [108-90-7]) and 3 sensitizers predicted as non-sensitizers (butyl glycidyl ether [2426-08-6], aniline [62-53-3], formaldehyde [50-00-0]). For vanillin [121-33-5], 15 similar compounds were found in the dataset with varying sensitization, possibly explaining the misprediction. Chlorobenzene [108-90-7] and butyl glycidyl ether [2426-08-6] had no similar analogues in the training set. Notably, the latter compound contained sub-structural fragments found as whole molecules within the dataset: 1-butanol [71-36-3], isopropanol [67-63-0] and propylene glycol [57-55-6], all labeled BMDC non-sensitizers, contrary to the sensitizer butyl glycidyl ether [2426-08-6]. The remaining 3 compounds (nonanoic acid [112-05-0], aniline [62-53-3], formaldehyde [50-00-0]) had similar analogues in the dataset, however, with varying sensitization labels. To improve the dataset's performance, the importance of chemical diversity and the presence of specific chemical series were acknowledged. Therefore, to enhance the model's accuracy, the plan is to expand the dataset by adding new chemical structures and conducting further SAR analysis.

The model was implemented in ISIDA-Predictor web-service⁵, which is freely accessible. By selecting the “Activity” option and then “Skin sensitization (BMDC) - Classification” from the “General kind of property” section, users can predict the sensitization potential of molecules by drawing them or uploading molecular structures in MDL SDF (Structured Data File) format. Screenshots of the ISIDA Predictor showing the input and output are provided in Figure 4.

The developed QSAR consensus model holds great promise in predicting the sensitization potential of chemicals using the BMDC assay. Its use of a large and consistent dataset aligns with the principles of the 3Rs (Replace, Reduce, Refine) and reduces the need of animal testing. This model

⁵The web-service of the Laboratory of Chemoinformatics: <https://chematlas.chimie.unistra.fr/cgi-bin/predictor2.cgi> (accessed 31 August 2023), “Skin sensitization (BMDC) - Classification” model in the “Activity” section.

enables rapid screening of chemical libraries, assessing toxicity for numerous compounds, including those not yet synthesized. Moreover, it can be a valuable tool when the *in vitro* BMDC test is not feasible due to certain physicochemical properties of the compound or regulatory restrictions.



Predictor

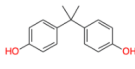
Selected model : Skin_sensitization_BMDC_2CIs

Please note that only the first 100 molecules have been sent to the Predictor.

All calculations processed!

Thank you for your patience.

[Explore results with the scrollbar](#)

Molecule Id	Molecule Name	Predicted value	Applied models	Prediction confidence	2D structure	Comments
1		Sensitizer	11/13	●		Good prediction confidence : AD satisfied for 85% of individual models and repartition of classes=63.64%.

[Back to Main Menu](#) [Download results](#)

Figure 4. Screenshots showing example of request for ISIDA (In Silico Design and Data Analysis) Predictor web service. Image A shows the ISIDA Predictor configuration page, where a user can select “Activity” general kind of property and then choose “Skin sensitization (BMDC) - Classification” from the list of models. Image B illustrates an output of ISIDA Predictor. Color code of prediction confidence is as follows: green – optimal; blue – good; orange – average; red – unreliable.

The predictive capabilities of this QSAR model hold significant value in compound prioritization, optimization, and risk assessment. Its recognition and acceptance by regulatory agencies such as the U.S. Environmental Protection Agency (EPA) and the European Chemicals Agency (ECHA) could reinforce its potential as a useful tool in chemical safety assessment. By meeting regulatory requirements for predicting sensitization endpoints, the model reduces the reliance on experimental assays for regulatory purposes, aiding industries in compiling a REACH dossier. Overall, this *in silico* QSAR model for predicting the sensitization potential of chemical compounds based on BMDC test data has achieved its objective, offering a valuable tool in toxicological and regulatory assessments.

4. Conclusion

In conclusion, the BMDC assay demonstrated strong predictive performance compared to other *in vitro* and *in chemico* assays with respect to LLNA and human labels, making it one of the best models for predicting KE3 output. The effectiveness of the assay is further highlighted by its exclusive correct predictions for certain compound. The first QSAR model trained on BMDC data to predict sensitization output achieved good performance during external cross-validation ($BA_{s-cv} = 0.82$). However, the size of the dataset, already substantial with 117 compounds (and destined to grow over time), limits its applicability domain. Nonetheless, this publicly accessible⁶ QSAR model can assist experts in making preliminary assessments of the sensitization potential of compounds of interest. With both *in vitro* and *in silico* tools based on BMDC model, the scientific community can comprehensively evaluate the sensitizing potential of chemical substances. The model's ability to prioritize substances for experimental testing and detect mismatches between prediction and experiment is valuable for early detection of sensitization potential.

5. References

1. OECD. The Adverse Outcome Pathway for Skin Sensitisation Initiated by Covalent Binding to Proteins. Published online 2014. <https://www.oecd-ilibrary.org/content/publication/9789264221444-en>
2. Sullivan KM, Enoch SJ, Ezendam J, Sewald K, Roggen EL, Cochrane S. An Adverse Outcome Pathway for Sensitization of the Respiratory Tract by Low-Molecular-Weight Chemicals: Building Evidence to Support the Utility of *In Vitro* and *In Silico* Methods in a Regulatory Context. *Appl Vitro Toxicol*. 2017;3(3):213-226. doi:10.1089/aivt.2017.0010
3. European Parliament and Council. *Regulation (EC) No 1223/2009 of the European Parliament and of the Council of 30 November 2009 on Cosmetic Products (Recast) (Text with EEA Relevance)*. Vol 342.; 2009. Accessed October 27, 2022. <http://data.europa.eu/eli/reg/2009/1223/oj/eng>
4. OCDE. Test No. 442E: In Vitro Skin Sensitisation. Published online 2022. <https://www.oecd-ilibrary.org/content/publication/9789264264359-en>
5. OCDE. Test No. 442C: In Chemico Skin Sensitisation. Published online 2022. <https://www.oecd-ilibrary.org/content/publication/9789264229709-en>
6. OCDE. Test No. 442D: In Vitro Skin Sensitisation. Published online 2022. <https://www.oecd-ilibrary.org/content/publication/9789264229822-en>
7. Kimber I, Dearman RJ, Scholes EW, Basketter DA. The local lymph node assay: developments and applications. *Toxicology*. 1994;93(1):13-31. doi:10.1016/0300-483X(94)90193-7
8. OCDE. Guideline No. 497: Defined Approaches on Skin Sensitisation. Published online 2021. <https://www.oecd-ilibrary.org/content/publication/b92879a4-en>
9. Battais F, Langonne I, Muller S, et al. The BMDC model, a performant cell-based test to assess the sensitising potential and potency of chemicals including pre/pro-haptens. *Contact Dermatitis*. 2023;under revision.
10. Borba JVB, Braga RC, Alves VM, et al. Pred-Skin: A Web Portal for Accurate Prediction of Human Skin Sensitizers. *Chem Res Toxicol*. 2021;34(2):258-267. doi:10.1021/acs.chemrestox.0c00186
11. Berthold MR, Cebron N, Dill F, et al. KNIME: The Konstanz Information Miner. In: *Studies in Classification, Data Analysis, and Knowledge Organization (GfKL 2007)*. Springer; 2007.
12. Chang CC, Lin CJ. LIBSVM: A library for support vector machines. *ACM Trans Intell Syst Technol*. 2011;2(3):27:1-27:27.

⁶ The web-service of the Laboratory of Chemoinformatics: <https://chematlas.chimie.unistra.fr/cgi-bin/predictor2.cgi> (accessed 31 August 2023), "Skin sensitization (BMDC) - Classification" model in the "Activity" section.

13. Horvath D, Marcou G, Varnek A. A unified approach to the applicability domain problem of QSAR models. *J Cheminformatics*. 2010;2(S1):O6. doi:10.1186/1758-2946-2-S1-O6
14. Marzulli FN, Maibach HI. Contact allergy: Predictive testing in man. *Contact Dermatitis*. 1976;2(1):1-17. doi:10.1111/j.1600-0536.1976.tb02972.x
15. Basketter DA, Scholes EW, Wahlkvist H, Montelius J. An evaluation of the suitability of benzocaine as a positive control skin sensitizer. *Contact Dermatitis*. 1995;33(1):28-32. doi:10.1111/j.1600-0536.1995.tb00443.x
16. Basketter DA, Clapp C, Jefferies D, et al. Predictive identification of human skin sensitization thresholds. *Contact Dermatitis*. 2005;53(5):260-267. doi:10.1111/j.0105-1873.2005.00707.x
17. Basketter D, White IR, McFadden JP, Kimber I. Hexyl cinnamal: consideration of skin-sensitizing properties and suitability as a positive control. *Cutan Ocul Toxicol*. 2015;34(3):227-231. doi:10.3109/15569527.2014.933973
18. Rustemeyer T, de Groot J, von Blomberg BME, Frosch PJ, Scheper RJ. Cross-Reactivity Patterns of Contact-Sensitizing Methacrylates. *Toxicol Appl Pharmacol*. 1998;148(1):83-90. doi:10.1006/taap.1997.8304
19. Schneider K, Akkan Z. Quantitative relationship between the local lymph node assay and human skin sensitization assays. *Regul Toxicol Pharmacol*. 2004;39(3):245-255. doi:10.1016/j.yrtph.2004.02.002
20. Vocanson M, Goujon C, Chabeau G, et al. The skin allergenic properties of chemicals may depend on contaminants—evidence from studies on coumarin. *Int Arch Allergy Immunol*. 2006;140(3):231-238. doi:10.1159/000093248
21. Roberts DW, Api AM. Chemical applicability domain of the local lymph node assay (LLNA) for skin sensitisation potency. Part 4. Quantitative correlation of LLNA potency with human potency. *Regul Toxicol Pharmacol*. 2018;96:76-84. doi:10.1016/j.yrtph.2018.04.022
22. Basketter DA, Scholes EW, Kimber I. The performance of the local lymph node assay with chemicals identified as contact allergens in the human maximization test. *Food Chem Toxicol*. 1994;32(6):543-547. doi:10.1016/0278-6915(94)90112-0
23. Bauch C, Kolle SN, Ramirez T, et al. Putting the parts together: Combining in vitro methods to test for skin sensitizing potentials. *Regul Toxicol Pharmacol*. 2012;63(3):489-504. doi:10.1016/j.yrtph.2012.05.013
24. Assaf Vandecasteele H, Gautier F, Tourneix F, Vliet E van, Bury D, Alépée N. Next generation risk assessment for skin sensitisation: A case study with propyl paraben. *Regul Toxicol Pharmacol*. 2021;123:104936. doi:10.1016/j.yrtph.2021.104936
25. Roberts DW, Aptula AO, Patlewicz G. Electrophilic Chemistry Related to Skin Sensitization. Reaction Mechanistic Applicability Domain Classification for a Published Data Set of 106 Chemicals Tested in the Mouse Local Lymph Node Assay. *Chem Res Toxicol*. 2007;20(1):44-60. doi:10.1021/tx060121y
26. Urbisch D, Mehling A, Guth K, et al. Assessing skin sensitization hazard in mice and men using non-animal test methods. *Regul Toxicol Pharmacol*. 2015;71(2):337-351. doi:10.1016/j.yrtph.2014.12.008

Supporting Information of “Benchmarking and QSAR Modeling of BMDC Assay for Identifying Sensitizing Chemicals”

Lisa Chedik^{1#} and Shamkhal Baybekov^{2#}, Gilles Marcou², Frédéric Cosnier¹, Mélanie Mourot-Bousquenaud¹, Sandrine Jacquenet¹, Alexandre Varnek², Fabrice Battais¹

¹ Institut national de recherche et de sécurité pour la prévention des accidents du travail et des maladies professionnelles (INRS), Dept Toxicologie et Biométrie, 1 rue du Morvan, 54519 Vandoeuvre-lès-Nancy, France

² Laboratory of Chemoinformatics UMR 7140 CNRS, University of Strasbourg, Strasbourg, France

These authors contributed equally to this work.

Supplemental 1. Sensitization labels in human and according to different tests (S: sensitizer; NS: non sensitizer)

CAS	Compound	Human	LLNA	BMDC	DPRA	KeratinoSens	LuSens	U-SENS	hCLAT	mMUSST	PredSkin
2426-08-6	Butyl glycidyl ether	S	S	S	S	S	S	S	NS	S	S
4903-09-7	3-Chloro-4-methoxybenzaldehyde		NS	S	S	NS	S	S			NS
100-06-1	Methoxyacetophenone	NS	NS	NS	NS	S	S	S		NS	NS
100-11-8	4-Nitrobenzyl bromide		S	S	S	S	S	NS	S	S	S
100-43-6	4-Vinyl pyridine		S	S	S	S		S			S
100-52-7	Benzaldehyde	NS	NS	NS	NS	S		S			NS
101-39-3	a-methyl-trans-cinnamaldehyde	NS	S	S	S	S		S	S		NS
101-86-0	Hexyl cinnamic aldehyde	NS	S	S	NS		S	S	NS	NS	NS
103-11-7	2-Ethylhexyl acrylate		S	S	S	S	S	NS	S	S	S
103-95-7	Cyclamen aldehyde	NS	S	S							NS
104-27-8	1,4-Methoxyphenyl-1-penten-3-one		S	S	S	S		S			S
104-54-1	Cinnamyl Alcohol	S	S	S	S	S	S	S	S	NS	S
104-55-2	Cinnamic aldehyde	S	S	S	S	S	S	S	S	S	S
106-24-1	Geraniol	S	S	S	NS	S		S	S		S
106-50-3	1,4-Phenylenediamine	S	S	S	S	S	S	S	S	S	S
106-51-4	p-benzoquinone	S	S	S	S	S	S	S	S	S	S
107-15-3	Ethylenediamine free base	S	S	S	NS	S	NS	S	S	S	S
107-22-2	Glyoxal	S	S	S	S	S	S	S	S	S	S
107-75-5	Hydroxycitronellal	S	S	S	S	S	S	S	S	S	S
108-31-6	Maleic anhydride	S	S	S	S	S		S	S		S
108-46-3	Resorcinol	S	S	S		NS	NS	S	S		S
108-90-7	Chlorobenzene	NS	NS	NS	NS	NS		NS			NS

2

CAS	Compound	Human	LLNA	BMDC	DPRA	KeratinoSens	LuSens	U-SENS	hCLAT	mMUSST	PredSkin
109-55-7	3-Dimethylamino-1-propylamine	S	S	S	NS	S			S		S
109-65-9	1-bromobutane		NS	NS		NS		NS	S		NS
110-54-3	Hexane		NS	NS	NS		NS	NS	NS	NS	NS
111-25-1	1-bromohexane		S	S		S			NS	NS	NS
111-30-8	Glutaraldehyde	S	S	S	S	S	S	S	S		S
111-80-8	Methyl 2 nonynoate	S	S	S	S	S			S		S
1118-71-4	2,2,6,6-tetramethyl-3,5-heptanedione		S	S	NS	NS		NS			S
112-05-0	Nonanoic acid		NS	NS	NS	NS		S	S		NS
1154-59-2	Tetrachloro-salicylanilide	S	S	S	S	S			S		S
116-26-7	Safranal		S	S	S	S		S			NS
1166-52-5	Lauryl gallate	S	S	S	S	S			S		S
119-36-8	Methyl salicylate	NS	NS	NS	NS	NS	S	NS	NS	NS	NS
119-84-6	Dihydro Coumarin	S	S	NS	S	NS		NS	S		S
120-51-4	Benzyl benzoate	NS	S	NS	NS	S			NS	NS	NS
121-32-4	Ethyl vanillin	NS	NS	S		S		NS	NS		NS
121-33-5	Vanilline	NS	NS	NS			NS	S	NS	NS	NS
121-57-3	Sulfanilic acid	NS	NS	NS	NS	NS			NS	NS	NS
121-79-9	Propyl gallate	S	S	S	S	S	NS	S	S	S	S
122-40-7	a-amylcinnamaldehyde	S	S	S	NS			S	S		S
122-57-6	Benzylideneacetone	S	S	S	S	S	S	S	S	S	S
122-78-1	Phenylacetaldehyde	S	S	S	S	S			S		S
123-31-9	Hydroquinone	S	S	S	S	S		S	S		S
124-07-2	Octanoic acid	NS	NS	NS	NS	NS		S	S		NS
124-12-9	heptyl cyanide		NS	NS	NS	NS		NS			NS

3

CAS	Compound	Human	LLNA	BMDC	DPRA	KeratinoSens	LuSens	U-SENS	hCLAT	mMUSST	PredSkin
13706-86-0	5-Methyl-2,3-hexanedione	S	S	S	S	S		S	S		NS
137-26-8	Tetramethylthiuram disulfide	S	S	S	S	S	S	S	S		S
138-89-6	N, N-dimethyl-4-nitrosaniline		S	S	S	S		S			NS
140-67-0	4-allylanisole		S	S	S		NS	NS	S	S	S
141-05-9	Diethyl maleate	S	S	S							S
149-30-4	2 Mercaptobenzothiasole	S	S	S	S	S	S	S	S	S	S
151-21-3	Sodium dodecyl sulfate		S	NS		NS	NS		NS	NS	NS
15646-46-5	Oxazolone	S	S	S							S
1675-54-3	Bisphenol A-diglycidyl ether	S	S	S	S	S		S	S		S
20048-27-5	bandrowski's base	S	S	S							S
2111-75-3	Perillaldehyde	S	S	S	S	S		S	S		S
2277-19-2	cis-6-nonenal		S	S							S
2345-34-8	4 Acetoxybenzoic acid		S	NS	S	NS		NS			NS
25646-71-3	4-N-Ethyl-N-2-methan-sulphonamidoethyl-2-methyl-1,4-phenylenediamine		S	S							S
2634-33-5	1,2-benzisothiazol-3-2H-one	S	S	S	S	S		S	S		S
2682-20-4	2-Methyl-4-isothiazolin-3-one	S	S	NS	S	S		S	S		S
27813-02-1	2-hydroxypropyl methacrylate	S	NS	S	S	S		S	NS		S
2785-87-7	Dihydroeugenol		S	S	NS	S					NS
2835-95-2	5-amino-2-methylphenol		S	S	S	S		S			S
2835-99-6	4-Amino-m-cresol		S	S	S	S		S			S
2892-51-5	Squaric acid		S	NS	S	NS		NS			S
31906-04-4	Lyril	S	S	S	S	S		S	S		S
35691-65-7	1,2-Dibromo-2,4-dicyanobutane	S	S	S	S	S	S		S	NS	S
39236-46-9	Imidazolidinyl urea	S	S	S	S	S	S	S	S	S	S

4

5.2 Skin sensitization

CAS	Compound	Human	LLNA	BMDC	DPRA	KeratinoSens	LuSens	U-SENS	hCLAT	mMUSST	PredSkin
488-17-5	3-Methylcatechol		S	S	S	S		S			NS
50-00-0	Formaldehyde	S	S	S	S	S	S	S	S	S	S
50-21-5	Lactic acid	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS
514-10-3	Abietic acid	S	S	NS	S	S		S	NS		S
5231-87-8	Squaric acid diethyl ester		S	S	NS	NS		NS			S
5307-14-2	2-Nitro-1,4-phenyldiamine	S	S	S	S	S		S	S		S
5392-40-5	Citral	S	S	S	S	S	S	S	S	NS	S
56-81-5	Glycerol	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS
57-55-6	Propylene glycol	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS
579-07-7	1-Phenyl-1,2-propanedione		S	S	S	S		S	S		NS
591-27-5	3 Aminophenol	S	S	S	NS	NS		S	S		S
615-50-9	2,5-Diaminotoluene sulphate	S	S	S	S	S		S	S		S
624-49-7	Dimethylfumarate		S	S		S		S	S		NS
62-53-3	Aniline	S	S	S	NS		NS	S	S	S	S
63-74-1	Sulfanilamide	S	NS	NS	NS	NS	NS	NS	NS	NS	S
638-45-9	1-Iodohexane		NS	NS	S	S		NS	S		NS
65-85-0	Acide benzoique	NS	NS	NS	S	NS	NS	NS	NS	NS	NS
66-27-3	Methyl methanesulfonate		S	S	S	S		S	NS		NS
6728-26-3	trans-2-Hexenal	S	S	S	S	S			S		S
67-63-0	Isopropanol	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS
68-12-2	dimethylformamide		NS	NS	NS	NS			NS		S
69-72-7	Acide salicylique	NS	NS	NS	NS	NS	NS	NS	S	NS	NS
70-34-8	dinitrofluorobenzene		S	S		S			S		S
71-36-3	1-Butanol	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS

5

CAS	Compound	Human	LLNA	BMDC	DPRA	KeratinoSens	LuSens	U-SENS	hCLAT	mMUSST	PredSkin
78-70-6	Linalool	S	S	S	NS	NS		S	S		S
80-54-6	Lilial	S	S	S		NS		S	S		NS
81-07-2	Saccharine		NS	NS		NS		NS	NS		S
818-61-1	2-hydroxyethyl acrylate	S	S	S	S	S		S	S		S
84-66-2	Diethyl phthalate	NS	NS	NS	NS	NS	NS		S		NS
874-23-7	2-Acetyl-cyclohexanone		NS	NS	S	S		S	S		S
87-86-5	Pentachlorophenol	NS	S	S	S	NS			S		NS
885-62-1	2,4-Dinitrobenzenesulfonic acid,sodium salt		S	S							S
886-38-4	Diphenylcyclopropanone	S	S	S	S	S		S	S		S
90-15-3	1-Naphtol		S	S	S	S		S	S		S
91-64-5	Coumarin	S	NS	S	NS	S			NS		S
92-48-8	6-Methylcoumarin	NS	NS	S	NS	S	S	NS		NS	NS
93-51-6	2-methoxy-4-methylphenol	NS	S	S		NS		S	S		NS
93-53-8	2-Phenylpropionaldehyde		S	S	S	S	S	S	S	NS	NS
94-09-7	Benzocaine	S	NS	NS		S		S	S		S
94-13-3	Propylparaben	NS	NS	S	NS	S	S	S	S	NS	NS
94-36-0	Benzoyl peroxide	S	S	S	S	NS	NS	S	NS		S
95-55-6	2 Aminophenol	S	S	S	S	S		S	S		S
97-53-0	Eugenol		S	S	S	NS	S	S	S	S	S
97-54-1	Isoeugenol	S	S	S	S	S	S	S	NS	NS	S
97-90-5	Ethylene glycol dimethacrylate	S	S	S	S	S	S	S	S	S	S
97-96-1	Diethyl acetaldehyde		S	NS	S	S		NS	S		NS
99-76-3	methyl paraben		NS	NS	NS	S	S	S			S
99-96-7	4 hydroxy benzoic acid		NS	NS	NS	NS	NS	NS		NS	NS

6

Supplemental 2. QSAR modeling

Input data in the QSAR model was standardized for compatibility and comparability by removing stereochemistry, dearomatization or by handling the tautomeric forms. Inorganic compounds (salts) were also omitted due to limitations in applicable descriptors for organic-focused QSAR. Furthermore, three complex mixtures were excluded from the dataset because their exact compositions were undisclosed, hindering the attribution of biological activity.

The dataset was divided into 5 sets for training (blue) and external testing (red) (**Figure S2.1**). A 5-fold external cross-validation was performed, using the training sets for model building and validation. The test sets were used solely to evaluate the external cross-validation performances of the final models. No decisions were made based on the external cross-validation performance measures.

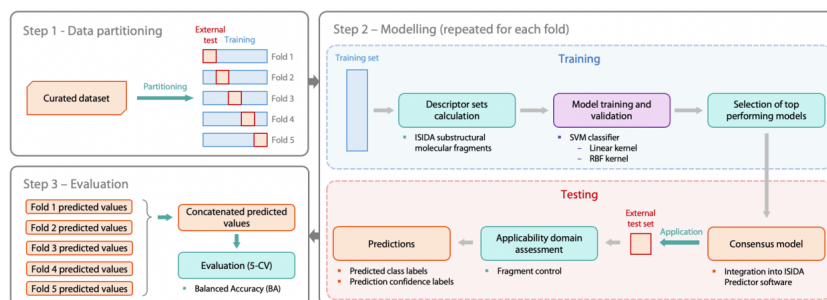


Figure S2.1. Workflow of the 5-fold external cross-validation (5-CV). The external test sets are strictly isolated from the whole modelling, validation and publication procedure. Molecular descriptors, machine learning optimization and model selection are performed via internal cross-validation and training and tuning sets split procedures.

For each fold's training set, 115 sets of ISIDA sub-structural molecular fragment descriptors (atom-centered fragments, sequences of atoms and bonds, triplets)⁹ were calculated. Both linear and radial basis function (RBF) kernel Support Vector Machine (SVM) classification models were built for each descriptor set. This study utilized a range of fragmentation types, which are listed in **Table S2.1**. Fragment length was set to a minimum ranging between 2 and 3, and a maximum ranging between 3 and 5. During the fragmentation process, additional options such as "Atom Pairs" and "Do All Ways" were utilized. "Atom pairs" removes constitutional details and only provides the number of constitutive atoms, while "Do All Ways" searches for all paths connecting two atoms in sequence fragments.

Table S2.1. Description of fragmentation types used in this study and their notations.

Fragmentation type	Notation
Sequences of atoms only	IA
Sequences of bonds only	IB
Sequences of atoms and bonds	IAB
Atom centered fragments based on sequences of atoms	IIA
Atom centered fragments based on sequences of bonds	IIB
Atom centered fragments based on sequences of atoms and bonds	IIAB
Atom centered fragments based on sequences of atoms of fixed length	IIA_R
Atom centered fragments based on sequences of atoms and bonds of fixed length	IIAB_R
Triplets	IIIA

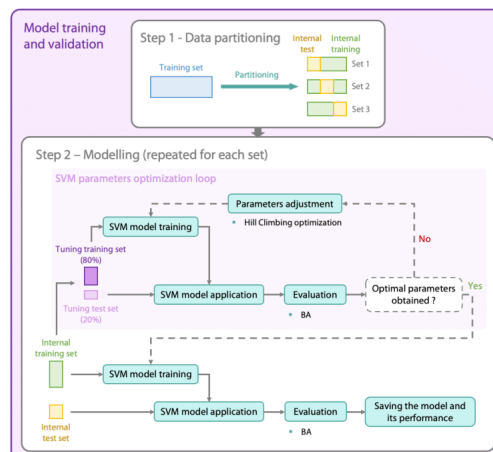


Figure S2.2. Workflow of the model training and validation. It involves partitioning of the dataset into smaller sets of internal training (green) and internal test (yellow) sets. Each internal training set is divided into 2 subsets that will be used for finding optimal set of SVM hyperparameters. For each internal training set a model is then trained using optimum parameters, followed by the evaluation on the corresponding internal test set.

The model training process is illustrated in **Figure S2.2**. The initial training set (blue) is split into 3 sets for internal training (green) and internal testing (yellow). Models are built on internal training sets and validated on internal test sets. SVM hyperparameters are optimized separately for each internal training set using an 80% (tuning train) - 20% (tuning test) partition (in dark purple and light purple, respectively). The Hill climbing method is used to maximize observed balanced accuracy on the tuning test set. Optimal hyperparameters were used to train an SVM model on the complete internal training set (in green on **Figure S2.2**), followed by the application of the model on the internal test set. The procedure resulted in 3 optimized SVM classification models and their internal cross-validation performances, for each descriptor set.

Best models (internal cross-validated BA larger than 0.9) entered a consensus model. A list of the 45 individual models that were selected for the consensus model is provided in **Table S2.2**. The consensus model was then integrated into ISIDA Predictor software and applied to the external test set (red) of the corresponding fold. The predicted values on all external validation sets were concatenated and the external cross-validation performances reported (**Table 3**).

Table S2.2. List of SVM models constituting the consensus model. Information about used descriptor spaces and internal validation BA are given. “P” – “Atom Pairs”, “AP” – “Do All Ways”. Other notations are described in the previous section.

Descriptor space	BA
IB(3-5)_AP	1
IB(3-5)	1
IAB(2-4)_AP	1
IAB(2-4)_AP	1
IAB(2-4)_P	1
IAB(2-4)	1
IAB(2-5)_AP	1
IAB(2-5)_AP	1
IAB(2-5)	1
IAB(2-5)	1
IAB(3-5)_P	1
IIB(2-3)	1
IIB(3-3)	1
IIAB(3-4)	1
IIA(2-5)_R	1
IIA(3-3)_R	1
IIAB(3-4)_R_P	1
IIIA(2-4)	1
IB(2-4)_AP	0.944
IB(2-4)	0.944
IB(2-5)_AP	0.944
IB(3-5)	0.944
IAB(2-3)_P	0.944
IAB(2-4)_AP	0.944
IAB(2-4)	0.944
IAB(2-5)_P	0.944
IAB(3-3)_AP	0.944
IAB(3-4)_AP	0.944
IAB(3-4)_AP	0.944
IAB(3-4)	0.944
IAB(3-5)_AP	0.944
IIB(2-4)	0.944
IIA(2-4)_R	0.944
IIA(3-5)_R	0.944
IIB(2-4)_R	0.944
IB(3-4)	0.938
IB(3-5)_AP	0.938
IAB(3-5)_AP	0.938
IIA(3-4)	0.938
IIA(2-3)_R	0.938
IIAB(2-3)_R_P	0.938
IIAB(2-4)_R_P	0.938
IIAB(3-4)_R	0.938
IIAB(3-5)_R	0.938
IIAB(3-5)_R	0.938

Supplemental 3. Performances of skin sensitization assays targeting KE3 (Key Event 3) with respect to Local Lymph Node Assay (LLNA) labels. The evaluation was performed on a set of common compounds (n = 30).

Assay	mMUSST	h-CLAT	U-SENS TM	BMDC
Data source	ICE	ICE	ICE	Battais 2023
BA	0.87	0.86	0.86	0.96
Accuracy	0.83	0.87	0.87	0.97
Sensitivity	0.74	0.9	0.9	1
Specificity	1	0.82	0.82	0.91

Supplemental 4. Performances of skin sensitization assays targeting KE3 (Key Event 3) with respect to human labels. The evaluation was performed on a set of common compounds (n = 27).

Assay	mMUSST	h-CLAT	U-SENS TM	BMDC
Data source	ICE	ICE	ICE	Battais 2023
BA	0.88	0.85	0.83	0.88
Accuracy	0.85	0.85	0.85	0.89
Sensitivity	0.75	0.88	0.94	0.94
Specificity	1	0.82	0.73	0.82

5.2.2 Summary

The comparative study conducted in the article highlights the superior performance of the BMDC in skin sensitization assessment when compared to various *in vitro*, *in chemico*, and *in silico* assays. The LLNA and human sensitization potential of some compounds could be anticipated based on the BMDC assay results only. The developed QSAR model trained on BMDC data is additional proof of the consistency of the BMDC assay and offers significant assistance to experts. Predicting the BMDC output upstream of the actual experimental assay enables the prioritization of compound analysis thus reducing time, resource, and material expenditure. Moreover, integration of the BMDC assay and/or the QSAR model with other existing assays has the potential to establish a novel approach that effectively replaces *in vivo* tests, significantly improving the efficiency and reliability of skin sensitization evaluations for chemical substances.

5.3 Skin permeability

5.3.1 Introduction

The skin serves as a protective barrier against external agents, primarily due to the properties of its outermost layer, the stratum corneum (SC). However, the skin is not impervious, and xenobiotics can penetrate the SC, diffuse into the viable epidermis, and enter the general circulation through dermal capillaries.⁸⁶ The assessment of skin permeation is therefore crucial not only for the pharmaceutical and cosmetic industries but also for ensuring occupational safety where workers may be exposed to harmful substances during handling.¹⁰ While experimental validation remains essential, *in silico* methods have emerged as valuable tools for the preliminary evaluation of skin absorption.¹¹ The majority of the predictive models are linear equations built using physicochemical properties, such as lipophilicity and molecular weight^{87,88}, and are trained on scarce amount of data.^{11,89} The effectiveness of these models relies heavily on the availability of comprehensive and up-to-date training data.

To address this need, a collaborative effort with the researchers from the Institut National de la Recherche et de Sécurité (INRS) was undertaken, resulting in the meticulous compilation of a new dataset, called SkinPiX (Skin Permeation of identified Xenobiotics)¹³, comprising skin permeability data published between 2012 and 2021, thereby complementing the existing HuskinDB¹². The results of SkinPiX data collection and curation are available in the article provided in this chapter. The model trained on HuskinDB was applied to SkinPiX and chemical space coverage of both datasets was analyzed. The results of modelling the combination of HuskinDB and SkinPiX data provide a number of ideas on how to manage and analyze this data in order to improve QSAR models.

5.3.2 Data

There are two sources of data used in this work: HuskinDB and a new SkinPiX database compiled from literature published between 2012 and 2021. The HuskinDB database was used for training, whereas the SkinPiX database was used both for training and testing purposes. The description and curation of the SkinPiX database is provided in the article below.

SkinPiX

1

SkinPiX : An Update of Skin Permeability Data based on A Systematic Review of Recent Research

Abstract

The cutaneous absorption parameters of xenobiotics are crucial for the development of drugs and cosmetics, as well as for assessing environmental and occupational chemical risks. Despite the great variability in the design of experimental conditions due to uncertain international guidelines, datasets like HuskinDB have been created to report skin absorption endpoints. This review updates the available skin permeability data by rigorously compiling research published between 2012 and 2021. Inclusion and exclusion criteria have been selected to build the most harmonized and reusable dataset possible. The Generative Topographic Mapping method is applied to the present dataset and compared to HuskinDB to monitor the progress in skin permeability research and locate chemotypes of particular concern. The open-source dataset (SkinPiX) (available at <https://doi.org/10.57745/7FHQOY>) includes steady-state flux, maximum flux, lag time and permeability coefficient results for the substances tested, as well as relevant information on experimental parameters that can impact the data. It can be used to extract subsets of data for comparisons and to build predictive models.

Background & Summary

The skin plays an important protective role against external aggression, thanks mainly to the properties of its outermost layer: the *stratum corneum* (SC). However, the skin is not an absolute barrier and xenobiotics can penetrate the *stratum corneum*, diffuse into the viable epidermis and enter the general circulation through the capillaries of the dermis. The different steps of the transport process have been described elsewhere¹.

Accurate assessment of the rate and extent of the percutaneous absorption of xenobiotics is of paramount importance for the development of new pharmaceutical and cosmetic products applied to the skin to ensure or prevent their absorption into the deep layers of the skin. These data are also necessary to assess the chemical risk of substances when cutaneous environmental or occupational exposure exists.

These substances deposited on the skin can indeed be responsible for irritation, sensitizing effects or general toxic effects and require *ad hoc* regulatory labeling. For instance, the REACH Annex VII mentions skin sensitization, irritation and corrosion assessments for substances produced and imported into Europe in volumes above one ton. In addition, the dermal route of exposure must be addressed in Annex VI [Regulation (EC) No 1907/2006 of the European Parliament and of the Council of 18 December 2006 concerning the Registration, Evaluation, Authorization and Restriction of Chemicals (REACH), establishing a European Chemicals Agency, amending Directive 1999/45/EC and repealing Council Regulation (EEC) No 793/93 and Commission Regulation (EC) No 1488/94 as well as Council Directive 76/769/EEC and Commission Directives 91/155/EEC, 93/67/EEC, 93/105/EC and 2000/21/EC].

2

In practice, the permeation of a chemical substance through the skin, assimilated with a passive diffusion phenomenon, can be studied experimentally *in vitro* using a diffusion cell device composed of donor and acceptor compartments between which the skin (*stratum corneum* side up) is placed (Fig.1).

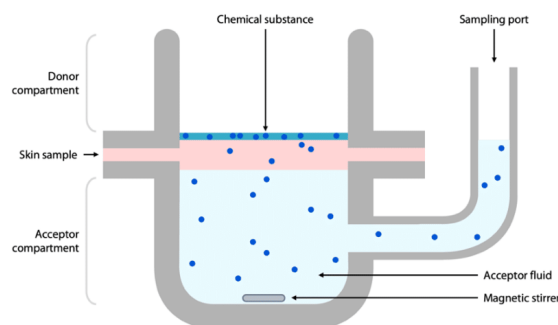


Figure 1. Schematic illustration of a static Franz diffusion cell.

These experiments measure the quantity of chemical (Q in μg) passing through the skin barrier per unit of skin surface (S in cm^2).

The experiments of percutaneous absorption can be conducted in finite dose conditions, i.e. a "finite" quantity of the chemical is applied to the skin so that a maximum flux (noted J_{peak}) of the test substance is achieved during a certain time interval (t_{peak}) but is not maintained (Fig.2). This contrasts with experiments with infinite doses where the concentration of the chemical in the donor compartment remains relatively constant throughout the experiment, ensuring the attainment and sustained maintenance of a steady-state flux J_{ss} .

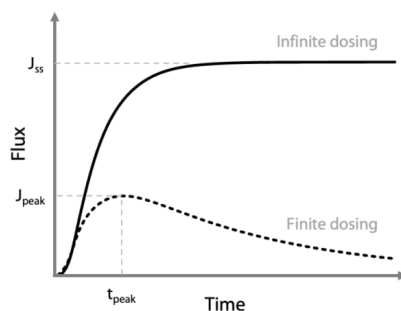


Figure 2. Theoretical change of outgoing flux for infinite (solid line) and finite dosing (dashed line). J_{peak} , t_{peak} , J_{ss} correspond to the maximum flux, the time of the maximum flux and the steady-state flux, respectively.

Using these infinite dose conditions and the steady-state flux data, it is possible to calculate the permeability coefficient, K_p with the following equation:

3

$$J_{ss} = K_p \times \Delta C_s$$

J_{ss} is the steady-state chemical transfer rate per unit area ($\mu\text{g}\cdot\text{cm}^{-2}\cdot\text{h}^{-1}$). Note that, when the substance is applied in pure form (neat liquid) or at its saturated concentration, the steady-state flux is called J_{max} .

ΔC_s is the difference in the concentration ($\mu\text{g}\cdot\text{cm}^{-3}$) of the chemical diffused between the inlet and outlet of the skin. Given the definition of infinite dose, ΔC_s is often approximated by the concentration of the chemical in the donor compartment at the beginning of the experiment (C_0).

K_p , the permeability coefficient ($\text{cm}\cdot\text{h}^{-1}$), reflects the ability of a membrane to let a substance permeate through it.

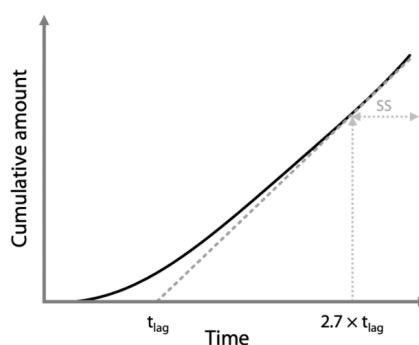


Figure 3. Cumulative amount of the tested chemical over time in the acceptor compartment during an infinite dose experiment. The solid line represents the whole experiment, and the dashed line represents the extrapolation of the linear steady-state phase (SS). The intersection of the dashed line with the time axis is the lag time (t_{lag}).

The amount of compound in the acceptor compartment increases exponentially over time until reaching the steady-state. J_{ss} is typically obtained from the slope of the linear part of the curve. The intersection of the linearized steady-state phase and time axis denotes the lag time, t_{lag} (Fig.3). The t_{lag} reflects the time it takes for the substance to cross the skin barrier.

Despite the fact that the first publications on the *in vitro* percutaneous absorption of xenobiotics date back to the 1960s, this research topic has not been studied extensively. Over the past 30 years, efforts have been made and initiatives taken to aggregate the available data on the skin permeation of xenobiotics. In 1990, Flynn collected human skin permeability coefficient data for the first time *in vitro* for over 90 chemicals². Then, the EDETOX database³ reported *in vivo* and *in vitro* literature data obtained for different species in a free databank which is still available on the web at <http://edetox.ncl.ac.uk> (updated in 2016). Samaras et al, extracted the *in vitro* human dataset from EDETOX and completed it with data obtained between 2001 and 2010⁴. This dataset is freely available for consultation only as a spreadsheet in the supplementary data. Finally, HuskinDB lists all the percutaneous absorption data from *in vitro* studies on human skin until 2011⁵. The corresponding database is freely accessible on <https://huskindb.drug-design.de> or

4

<https://doi.org/10.7303/syn21998881>, (last access 12/04/2023). Although this database represents a step forward compared to the two previous ones because it provides a better description of experimental conditions, it reports data on only 253 substances and as for the previous databases the inclusion/exclusion criteria conditions deserve to be more extensively described. It should be noted that in the publications selected for these different databases, not all the experimental conditions are systematically reported.

Cheruvu et al⁶ recently proposed an update to these data stemmed from a review paper⁷. The authors focused on maximal flux (J_{max}), and permeability coefficient (K_p) values collected from *in vitro* human skin permeation tests performed on human epidermal membranes or isolated stratum corneum at infinite dosing but the use of this latter type of skin can be debated (see usage notes, paragraph on skin layers). They also reported physicochemical properties and experimental conditions under which the data was generated (temperature, skin thickness, and skin integrity). Other parameters important for percutaneous absorption should have been reported (e.g. skin donor source, skin preparation techniques, skin source, storage duration and temperature, donor and acceptor pH, cell type).

The lack of data in the field of percutaneous absorption is particularly problematic for the generation of efficient predictive models on skin permeation such as QSPR (Quantitative Structure-Permeability Relationship) models. This implies that most existing *in silico* models are trained on the Flynn dataset² or variations of it, and have very limited domains of applicability.

In addition, the comparison of data between different publications can be tricky because, although international guidelines (OECD Guidance Document 28 (GD28) for conducting skin absorption studies⁸, Test Guideline 428 (TG428) for measuring skin absorption of chemicals *in vitro*⁹, and the OECD Guidance Notes 156 (GN156) on dermal absorption issued in 2019¹⁰ and 2022¹¹) give recommendations on experimental conditions and set-ups, they remain relatively imprecise and leave room for many variations in experimental designs that are left to the discretion of the experimenter. Many of these factors have a significant influence on the results of percutaneous absorption experiments¹², such as the donor type, also called vehicle^{13,14}, the skin donor type, the skin source site, the layer used and the experimental cell device (see usage notes).

Here we present SkinPiX (Skin Permeation of identified Xenobiotics), a new dataset obtained after the systematic collection of the available literature on human percutaneous absorption published after 2012. The dataset contains flux, t_{lag} and K_p data of the substances studied but also information specifying the experimental conditions. The scientific literature was curated manually by scientists from INRS (Reference body for occupational risk prevention in France), experts in percutaneous absorption. Exclusion or inclusion criteria were applied as explained in the Methods section.

When the information is available in the publication, SkinPiX indicates, in addition to the percutaneous absorption data, the experimental parameters in additional columns. HuskinDB was taken as a template, in order to facilitate the integration of these new data

5

into the database. Some columns have also been added compared to HuskinDB (Data ID, Publication ID, CAS number, Category donor type, Category acceptor type). The publication ID is the number assigned to each publication during the systematic literature search. An error column has been added for the following parameters: Permeability coefficient K_p , Steady-state Flux J_{ss} , Maximum Flux J_{max} , t_{lag} .

The influence of different experimental parameters is discussed further in this publication, so that the user of the dataset can choose a set of data consistent with another one regardless of the type of analysis it may need (for instance, QSPR modeling).

This set of reliable and harmonizable human percutaneous absorption data has been designed to serve as a reference for aggregate exposure and risk assessment by federal and state governments, universities, and for research and development on transdermal drug delivery by the pharmaceutical and cosmetics industries. Our belief is that this dataset has the potential to uncover commonly utilized experimental conditions, which could then be recommended in future versions of international guidelines. By harmonizing practices and reducing result variability, these guidelines would promote consistency and reliability across experiments.

This dataset is well-suited for data extraction and its quality and richness are also assets for the development of robust *in silico* models⁷.

Methods

We conducted a systematic literature search and scrupulously analyzed the publications of interest to obtain a comprehensive dataset. The general workflow for creating the dataset is shown in Figure 4. The aim was to cover as much as possible the new skin permeability data for well-defined organic compounds i.e. no UVCBs (unknown or variable composition, complex reaction products and biological materials) for instance, in an unambiguous experimental setup.

6

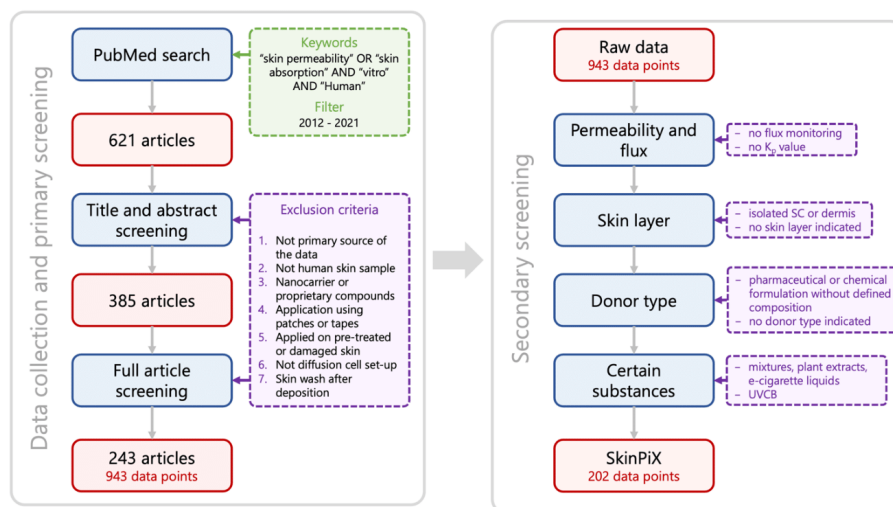


Figure 4. Data collection and filtering workflow. The process follows two main steps. First, relevant scientific publications were extracted using PubMed. Then skin permeability data were extracted along with relevant metadata. We kept only those data considered which met several criteria, as explained in section 1. "SC" stands for stratum corneum.

1. Inclusion and exclusion criteria

We performed a systematic and comprehensive literature review of percutaneous absorption existing *in vitro* experimental data obtained between 2012 and 2021, using an automated approach. We searched publications in the major electronic database: PubMed, with a date restriction from January 2012 to June 2021 and the keywords "skin permeability" OR "skin absorption" AND "vitro" AND "Human", resulting in 621 references in the public domain and the corresponding abstracts.

Considering the information in the abstracts, only publications in English and which were readily accessible were selected. The first manual sorting was performed and, according to the exclusion criteria for the analysis, we discarded publications:

- (1) without a primary source of the data (e.g., reviews, book chapters, datasets of published data or publications presenting predictive models, etc.) or already covered by the previous database HuskinDB (for 2012 publications);
- (2) with experiments conducted on animal, synthetic or artificial skin or Reconstructed Human Epidermis (RHE) and Skin (RHS) and Human Skin Equivalent (HSE) and other *in vitro* artificial skin models, cell lines and cultured skin, 3D organotypic constructs or experiments conducted *in vivo*;
- (3) with compounds formulated as nanocarriers, unidentified proprietary compounds;
- (4) with transdermal therapeutic systems (TTSs) of application of the substance, such as patches or tapes;

7

(5) with percutaneous absorption experiments performed on pre-treated skin (laser, microneedles, etc.) or damaged skin;

(6) without Franz cell or equivalent diffusion experimental set-up (microscopy, raman spectroscopy, microdialysis, transwell plate, etc.);

(7) with percutaneous absorption results impacted by a skin wash after deposition.

In all cases where the abstract did not allow verifying these criteria, or in case of doubt, the reference was kept.

At this stage, 385 references were retained. The full article was obtained in PDF form for each of them. The articles were read and those that included experiments/data that did not meet the previous seven exclusion criteria were excluded.

Experiments not reporting the flux monitoring of substances deposited on the skin (only fraction absorbed or quantity measured in the different skin layers/distribution) were discarded. If no K_p was mentioned and if it was not possible to calculate it from pK_p or concentration and J_{ss} , the reference was discarded.

Data obtained with isolated *stratum corneum* (SC) or isolated dermis alone were discarded (see usage notes) so that only data from experiments carried out on the epidermis or epidermis+dermis were selected. Given the lack of clarity in the guidelines on the use of full thickness skin (see usage notes), we included full thickness skin experiment data in the dataset. If skin layer was not mentioned, the publication was discarded. Due to the unclear recommendations of the guidelines on the use of epidermal membranes separated by the heat separation method (see usage notes), data obtained with epidermal membranes were kept in the dataset.

Similarly, due to the inconsistent guidelines regarding the determination of mass-balance recovery, which is defined as the percentage of the original substance recovered at the end of an experiment (refer to usage notes), and the absence of systematic reporting of this recovery in literature, we decided not to exclude data for which the reported recovery was poor (<80% or >120%), but when available, the recovery mentioned in the publication was reported in the column notes of the dataset.

As indicated in the usage notes section, the occlusion of the donor compartment may impact the percutaneous absorption parameters. We chose not to exclude data obtained with occlusion but to mention it in the column notes of SkinPiX dataset.

We chose not to exclude any data on the basis of the acceptor type mentioned. The data obtained with the deposit of neat substances were reported and a specific work on donor types was carried out. If we look at the number of counts per donor type from different publications, no single donor type really stands out, with a maximum of only 11 publications using water. Data for which the composition of the donor medium was not provided were excluded. Formulations with UVCB, such as MIGLYOL® 812 N, TWEEN® and poloxamer 407, were excluded. All pharmaceutical and chemical formulations of any kind were excluded

8

because the donor type was poorly identified, or the precise composition was not known, or their production over time was not guaranteed, and/or the composition could vary over time. Some vehicles contain known enhancers but we chose not to remove them from the dataset. However, when a publication studied specifically the effects of enhancers, only the results of the substance deposited in a donor type without enhancers were kept.

We chose not to exclude any data based on skin storage temperature and storage duration since these parameters were not reported in 20% to 30% of the endpoints analyzed. However, when a compound was tested on both fresh and frozen skins, we chose to keep only the fresh skin data.

We chose to keep data regardless of the experimental temperature reported because this information was not mentioned in more than 50% of the data points (for acceptor medium temperature). In the dataset, the reported temperatures range from 32°C to 37°C, but it was not always clear whether they corresponded to the donor compartment, the skin or the acceptor compartment. If a publication reported data for multiple experimental temperatures (within the framework of the study of a temperature effect), only data collected in the experiments closest to 32°C were kept.

Compounds meeting the definition of UVCB were excluded (compounds of vaping products and plant extracts). The parameters collected are indicated in the paragraph Data Records. Data processing was carried out using the KNIME Analytics Platform¹⁵. The KNIME workflows used to process the data are accessible in the online repository (<https://doi.org/10.57745/7FHQOY>) for transparency. The resulting SkinPiX dataset contained 202 data points.

2. Chemical space analysis of skin permeability data

The Generative Topographic Mapping (GTM) method¹⁶ was used to analyze the coverage of chemical space by HuskinDB and our new dataset. It is a dimensionality reduction method that transforms a multi-dimensional molecular descriptor space into a 2D latent space or a "map"¹⁷. This is accomplished by introducing a 2D manifold into the high-dimensional space and adjusting a normal probability density centered on it to fit the data distribution observed. Once the manifold is fitted, the compounds can be projected onto this 2D surface. The map can be colored based on population (density landscape) or property distribution (property/class landscape). The GTM class landscape was generated using ISIDA/GTM software.

Data records

For publications meeting all the inclusion criteria (see Methods, paragraph 1), the following information when available was collected and filled in an Excel sheet:

- Data ID and publication ID (integer): corresponds to the identifier given to each data entry and each unique publication. For a given publication ID, there can be several data ID with the same compound if there are percutaneous absorption experiments performed in different experimental conditions.
- SMILES (string): SMILES (Simplified Molecular Input Line Entry System) were extracted from the PubChem database, using the PubChem Identifier Exchange Service (<https://pubchem.ncbi.nlm.nih.gov/idxchange/idxchange.cgi>) by searching for molecules by their CAS (Chemical Abstracts Service) number.
- CAS number (string): unique and unambiguous CAS identifier that designates a specific substance. When not provided in the publication it was searched via PubChem (<https://pubchem.ncbi.nlm.nih.gov/>). CAS numbers of peptides are not defined.
- Compound name (string): for each substance, only one name or an amino acid sequence for peptide was entered in the dataset.
- K_p relation (string): signifies the exact K_p value ("=") or if K_p value is smaller (" $<$ ") or greater (" $>$ ") than the value given in the K_p column.
- K_p and K_p error in $\text{cm}\cdot\text{h}^{-1}$ (float): the permeability coefficient (K_p) value was obtained directly from the publication or calculated from the $\text{p}K_p$ ($\text{p}K_p = -\log K_p$) or calculated from J_{ss} and C_0 . The K_p (processed) column was used to harmonize K_p entries in decimal form. The same applied to $\log K_p$ (cm/s) (converted) column. When a range of K_p values has been reported in the publication, we have indicated the average K_p (processed) and $\log K_p$.
- Steady-state flux J_{ss} relation (string): signifies the exact J_{ss} value ("=") or if the J_{ss} value is smaller (" $<$ ") or greater (" $>$ ") than the value given in the steady-state flux J_{ss} column.
- Steady-state flux J_{ss} and J_{ss} error (float) were first reported as written in the publication with their original unit. Then steady-state flux J_{ss} (converted) and J_{ss} error (converted) in $\mu\text{g}\cdot\text{cm}^{-2}\cdot\text{h}^{-1}$ were also reported. These values were reported only if they were reported in the paper or if they could be calculated with the K_p and C_0 given. If necessary, conversions were performed in *ad hoc* units.
- Maximum flux J_{max} and maximum flux J_{max} error in $\mu\text{g}\cdot\text{cm}^{-2}\cdot\text{h}^{-1}$ (float): J_{max} was reported when the substance was dosed pure or in its saturation concentration. As for J_{ss} , J_{max} and its error were first reported as written in the publication with their original unit and were then reported in $\mu\text{g}\cdot\text{cm}^{-2}\cdot\text{h}^{-1}$.
- t_{lag} and t_{lag} error in h (float): t_{lag} and t_{lag} error were reported when the data were available. The column t_{lag} (h) (processed) harmonizes entries.
- Skin donor type (string): the human skin used was either from a cadaver or corresponded to discarded surgical skin.
- Skin source site (string): the anatomical area was indicated (abdomen, breast, back, thigh).
- Skin preparation (string): it corresponds to the treatment carried out on the full thickness skin to obtain the skin used for the experiments. But very often, experiments implement split thickness skins which have been dermatomed. The layers of skin can also be separated (heat separation) providing epidermal membranes.

10

- Layer used (string): this section specifies which skin layer(s) was (were) used for the experiment: epidermis alone, epidermis and dermis. Sometimes the layer used was not explicitly indicated but when the skin was dermatomed with a possible indication of the thickness, we could deduce the layer used.
- Storage duration (days) (integer): when the skin was used fresh, this box was filled with "0". In other cases, if the information was specified, then the storage duration was indicated in number of days or as a maximum number of days.
- Storage temperature (°C) (float): the skin was either used immediately or very quickly after collection (in this case, "used fresh" was indicated) or frozen or refrigerated before use. In these cases, the storage temperature was indicated.
- Donor type (string): indicated neat or diluted in a vehicle whose composition was given. Donor media were then classified into categories (column category donor type) (see Supporting materials).
- Donor pH (float): when provided, the pH value. If the experiments were carried out at different pH levels, only data relating to the pH levels most compatible with the skin were retained.
- Acceptor temperature (°C) and donor/skin surface temperature (°C) (float): the temperature was indicated if provided.
- Acceptor type (string): the composition of the acceptor medium was indicated. Acceptor media were then classified into categories (column category acceptor type) (see Supporting materials).
- Acceptor pH (float): when provided, the box was filled with the pH value.
- Cell type (string): type of permeation cell i.e. Franz diffusion cell (either static or flow through or modified Franz cell) or other type of diffusion cell. For Franz diffusion cells, if not specified in the publication, we have considered them to be static cells by default. If the publications did not explicitly mention that the experiments were carried out with Franz cells, we reported "other type of diffusion cell".
- Author (string): first author's name.
- Date of publication (integer): the year the article was published.
- DOI (string): DOI is an unambiguous identifier of scientific publications.
- Notes (string): The experts have provided information on whether the experiment was carried out in occlusive or semi-occlusive conditions, whether the authors of the publication have stated the use of infinite dose conditions or if the K_p value was derived from a finite dose scenario (see usage notes section). Additionally, the section also covers any indicated recovery process. It also highlights the calculation of parameters and provides any other relevant information for the reader's benefit.

If the parameter of interest was not mentioned in the source publication, it was annotated "N/A".

11

This new dataset aggregates 202 relevant endpoints (K_p) for 110 compounds of varying structures (drugs, industrial toxics, flame retardants, pesticides, etc.) from available literature data published since 2012.

This dataset allows filtering the data conveniently for each of the headings. One can for example seek all the lines of data which concern "benzoic acid". Several filters can be applied: e.g., CAS "58-22-0" (testosterone) and the skin donor type "cadaver".

The analysis of this dataset (Fig.5) provides several indications about the most frequent practices that could generally be accepted to build more robust guidance for measuring the K_p . For instance, the skin source is preferably back/thigh and dermatomed; the skin layers used are essentially composed of the epidermis and dermis; the acceptor and donor types are preferably PBS based formulations. Nevertheless, it is noteworthy that the common trend visible in Figure 5 is subject to the number of data points contributed by certain publications, particularly Ellison et al¹⁸ have the largest contribution to the dataset (47.5%).

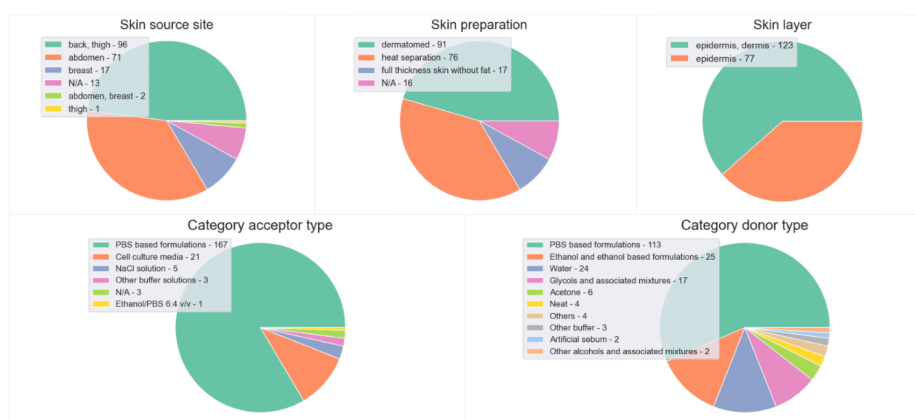


Figure 5. Overview of data distribution of SkinPiX dataset based on experimental parameters. The numbers in the legend correspond to the number of data points with respective labels.

Octanol-water partition coefficient ($\log P$) and molecular weight are substance-specific parameters that can have an impact on the permeability of the substance. For this reason, we have calculated $\log P$ and molecular weight using RDKit open-source cheminformatics software. $\log P$ is calculated using a method developed by Wildman and Crippen¹⁹, which takes into account contribution of each atom and its neighboring atoms to $\log P$. Figure 6 illustrates that the dataset contains diverse molecules in terms of calculated $\log P$, molecular weight and skin permeability coefficient. It contains chemicals exhibiting a diverse range of skin permeability, with $\log K_p$ values ranging from -6.2 to $0.26 \text{ cm}\cdot\text{h}^{-1}$. However, molecules with the highest count are relatively small (molecular weight = $[150; 200] \text{ g}\cdot\text{mol}^{-1}$), moderately lipophilic (calculated $\log P = [1; 2]$) and have moderate skin permeability ($\log K_p = [-6; -5]$). No specific relationship between skin permeability and calculated $\log P$ was established; however, compounds with high molecular weight ($> 500 \text{ g/mol}$) generally have low skin permeability, as expected for large molecules.

12

The compounds that were found most frequently in the dataset are caffeine (6 data points) and dichlorvos (6 data points).

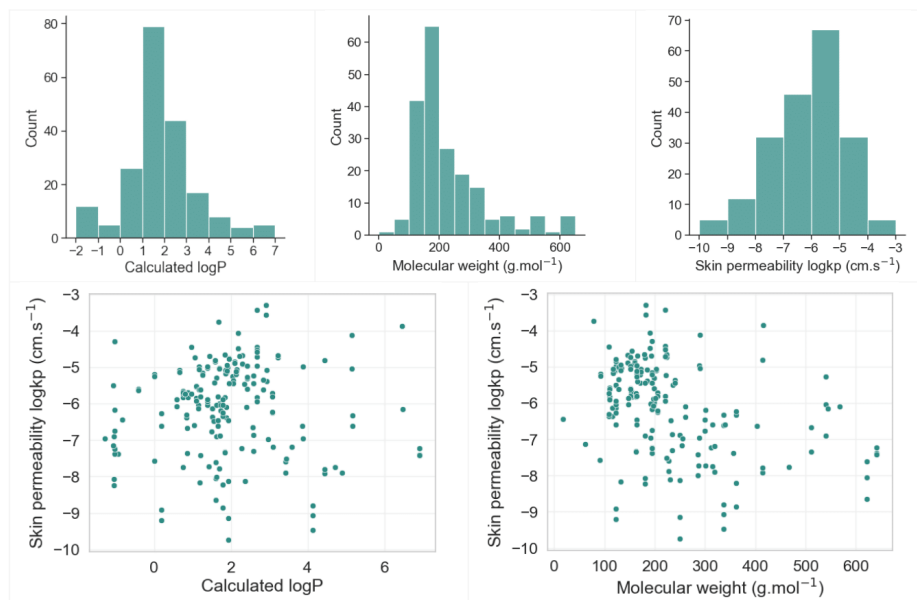


Figure 6. Overview of data distribution of SkinPiX dataset based on certain properties.

Chemical space analysis of skin permeability data

In this work, we applied GTM to visualize the chemical space coverage of skin permeability data by using HuskinDB and SkinPiX (Fig.7). The GTM class landscape shows that the main population of both datasets is located in the south-east quarter of the map (yellow, green and orange zones) while there are regions preferentially populated by HuskinDB (blue zones) or by SkinPiX (red zones). Examples of compounds and common chemical substructural features that are unique to SkinPiX (red zones) are indicated in Figure 7. However, the large part of SkinPiX covers the same region as HuskinDB: they are represented in the map as green to orange regions that are also the most densely populated. Most of these substructural features and compounds differ from their similar counterparts found in HuskinDB by additional or different structural decorations. The GTM analysis showed that SkinPiX expands the chemical space of skin permeability by introducing new molecular scaffolds.

13

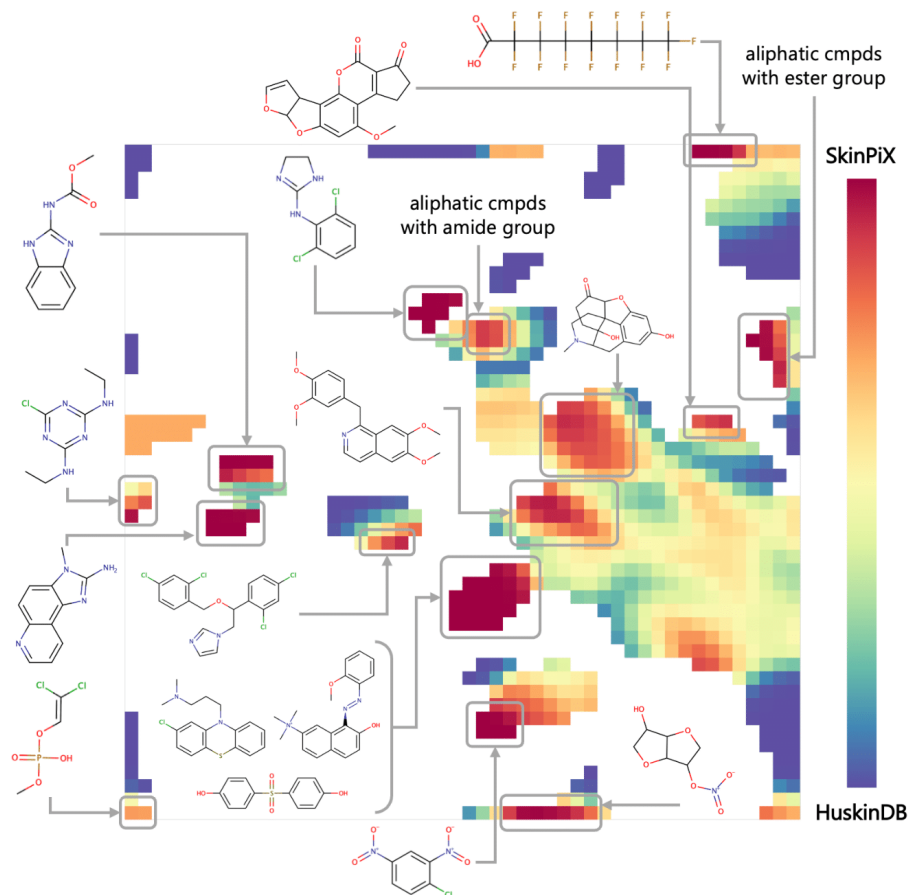


Figure 7. GTM landscape of skin permeability chemical space. Blue regions are mostly populated by compounds found in HuskinDB. Red regions are populated by compound data from SkinPiX. White regions do not contain any compound. The chemical content of various regions of the map is illustrated by example compounds (cmpds) and scaffolds.

Technical validation

The transcribed data were checked by a second reading by a researcher for accuracy and absence of error.

Usage Notes

One should be aware that the percutaneous absorption results of a given substance (K_p , J_{ss}) are influenced by many factors such as donor composition, acceptor composition and other experimental conditions^{12,20–22}.

14

Hence, when seeking to find the K_p or J_{ss} values for a particular substance from SkinPiX, it is beneficial to examine the experimental parameters that were utilized to obtain those results. Consulting the Usage notes section may help the user in determining whether the results for a given substance under different experimental conditions are comparable or not.

- Skin donor type

The guidelines proposed to carry out the studies either with human skin from autopsies (cadaver skin) or with surgical discard skin, the two main sources of supply¹⁰. Surgical discard skin is generally preferred but cadaver skin, which can be more convenient, is also accepted for percutaneous absorption studies as long as the integrity of the barrier is verified²¹, as the decrease of barrier integrity could lead to increased permeability. Nevertheless, as the skin is not considered viable, the metabolism of the substance cannot be studied. It is good practice to ensure before using cadaver skin that the skin does not metabolize the substance studied or that its metabolism does not have an impact on the flux. It is important to know how long and how the skin was kept before the experiment²¹. However, as the conditions in which cadaver skin is kept are variable, and the decomposition of the different components of post-mortem skin is a complex process, surgical samples are often preferred and recommended²³. Although the gender, age and phenotype of the skin donor may also have an impact on the percutaneous absorption of a substance²⁴, they were not taken into account in SkinPiX.

- Skin source site

Bormann et al, reviewed the impact of anatomical location on percutaneous penetration in humans *in vivo* with greater penetration on the face, neck and genital area²⁵. The differences observed *in vivo* can be explained by, among other things, the thickness of the SC, the density and size of hair follicles, hydration and the extent of blood irrigation^{12,24,26}. With *in vitro* percutaneous absorption experiments, a similar variability and the same trends have also been observed according to body zones^{27,28}.

The forearms and the hands are generally the most exposed cutaneous regions during occupational exposure to chemicals, but in practice most dermal absorption experiments use skin from abdomen or breast/chest skin samples obtained from aesthetic surgery, as mentioned in GN156 and GD28^{8,11}. Note that results sometimes include experiments performed on skins from several anatomical areas.

- The layer used

It is important to consider which layer of the skin was used to obtain experimental data when analyzing skin permeation values, as the different layers do not have the same permeability properties. Thinner skin thickness generally leads to a higher flow rate. But since the dermis is a predominantly hydrophilic tissue, its presence (in the case of dermatomed skin) or its absence (epidermis alone) has a greater impact on the percutaneous absorption of lipophilic substances²⁹. The latest version of GN156¹¹ recommends the use of split thickness skin of 200 to 400 μm which includes the SC, the viable epidermis and part of the dermis. The use of the viable epidermis and dermis in addition to the SC ensures better representation of the *in vivo* skin structure of the skin layers insofar as the viable epidermis and dermis can also have an impact on the diffusion of a chemical through the skin³⁰. Although they were proposed in GD28⁸ and TGD428⁹ dating from 2004, and in the 2019 version of GN156¹⁰, epidermal membranes (SC+ viable epidermis) obtained from heat separation no longer appear to be recommended in the latest GN156 version of 2022¹¹ insofar as they could, due to their insufficient barrier function, lead to overestimating the absorption results compared to dermatomed skin. The use of isolated SC presents a big disadvantage: this layer lower than 0.1mm is very fragile and it can be tricky to work on unaltered membranes with an intact barrier³¹. Note that there is no mention of the possible use of isolated SC in the current OECD guidelines.

15

It is clearly stated in the 2019 version of GN156 that full thickness skin cannot be used to determine flux¹⁰, certainly because the penetration of lipophilic substances is greatly reduced with full thickness skin compared to split thickness skin³². Surprisingly this information is not reported in the latest version of the guidance note¹¹.

- Skin preparation

In addition to skin used without preparation (full thickness skin without fat), either the skin is dermatomed to obtain split thickness skin of controlled thickness (see previous paragraph), or the epidermis is separated from the dermis. There are several epidermis-dermis separation methods³³, the most commonly used being heat separation. Epidermal membranes and dermatomed skin are both accepted even if, according to the 2019 version of GN156¹⁰, dermatomed skin is the most appropriate model. However, care must be taken to ensure that the heat separation technique does not alter the permeation properties of the skin. The method of skin preparation might have an impact on skin enzymes present in skin and can impact the results. In the case of esters, for example, the relevance of the data obtained with the epidermis is a subject of discussion as Lau et al, have shown that the heat separation technique could significantly decrease the activity of esterases³⁴.

- Skin storage temperature

For percutaneous absorption experiments, skin is generally used immediately after excision or at least within 24h (fresh skin) or stored frozen for up to several months according to GD28⁸. However, Dennerlein et al, questioning the validity of the experiments on which the guidance document is based, carried out experiments showing that up to 30 days of freezing at -20°C did not significantly alter the permeability of skin with respect to the 3 substances tested compared to freshly excised human skin³⁵. Jacques-Jamin et al, came to the same conclusion with 3 other substances and slightly longer freezing times of 8 and 12 weeks³⁶. On the other hand, storage at -80°C may increase permeability and is not recommended^{8,37}. For practical reasons, it is best to remove the subcutaneous tissue before freezing the skin. Repeated freezing and thawing are not recommended as this can damage the barrier. Frozen skin should not be used for substances metabolized by the skin, as the activity of enzymes may be altered and inactivated by freezing. Because the effect of freezing on the percutaneous absorption parameters of skin may depend on several factors, such as how the skin is frozen (full thickness, dermatomed, epidermal membranes), it is necessary to check the integrity of the barrier after storage in the freezer according to GD28⁸.

- Storage duration

The TGD428 recommends using fresh skin within 24 hours after excision⁹. Based on recent publications, if skin is stored frozen at -20°C, it must be kept for short periods of 1 to 3 months to obtain accurate and reliable permeation parameters^{35,36}.

- Cell type

Static, with appropriate continuous stirring of the acceptor fluid³⁸, and flow-through diffusion cells are both acceptable for skin *in vitro* absorption experiments according to all the OECD guidelines, insofar as they are composed of inert material^{8,9,11}. Some authors have summarized the advantages and drawbacks of each system^{23,39}.

Studies have shown similar results for these two types of cells^{40,41}. In the framework of their comparison study, Van de Sandt et al, concluded that the type of cell and its design have little impact on the results²⁹. The terms used to describe the cells used experimentally vary according to the authors, which does not always make it possible for them to be classified precisely.

- Experiment temperature

Numerous studies and GN156 indicate that experiment temperature is a crucial parameter to control as it affects the passive diffusion of substances and therefore their flux and lag time^{11,42-44}. That is why TGD428 and GD28 recommend keeping skin and the diffusion cell, in particular the acceptor chamber, at the physiological temperature of human skin, i.e. $32 \pm 1^\circ\text{C}$ ^{8,9}.

- Acceptor type

The type of acceptor is very important in *in vitro* percutaneous absorption experiments²¹. All the guidelines for *in vitro* dermal absorption testing agree that the type of acceptor used must not be a limiting factor in the permeation process⁴⁵. The solubility of the substance in the medium must be at least 10 times the maximum expected concentration (GN156)¹¹. The acceptor fluid should not affect skin integrity⁴⁵. GN156 proposes using a normal saline for hydrophilic substances and non-viable skin¹¹. For lipophilic compounds, GN156 indicates "the acceptor fluid may contain solvent mixtures such as ethanol and water (50% aqueous ethanol), <6% polyoxyethylene oleyl ether in water, or 5% bovine serum albumin"¹¹. However, in order to maintain viable skin, the acceptor should preferably be physiologically compatible with the skin (GD428 and GD28), such as a tissue culture medium, in particular to consider metabolism^{8,9}. An acceptor fluid with a high buffering capacity is required to guarantee the viability of the skin throughout the experiment. It is advisable to add glucose and antibiotics to the acceptor fluid to prevent the skin from deteriorating, especially for experiments lasting more than 24 hours⁴⁶. Its precise composition must be indicated.

Since the acceptor fluid has a major effect on skin absorption parameters, the guidelines should be more precise on this subject, as requested by a group of experts in the field²⁰, and should propose for each situation precise compositions of the acceptor fluids, which would ensure that the future K_p , J_{ss} and J_{max} data found in the literature are not impacted by this parameter.

Figure 5 shows the acceptor category types included in the dataset.

- Acceptor pH

Only GD28 gives information on the pH of the acceptor medium: "for non-viable skin preparations, the acceptor fluids for evaluating water soluble compounds are usually saline solutions, pH 7.4", which correspond to quite specific conditions⁸. The pH must take into account the more general recommendations on the acceptor type: it must not affect the integrity of the barrier, and adequate solubility of the test substance in the acceptor fluid should be demonstrated (TGD 428)⁹. As a general rule, the acceptor fluid is aqueous. Wagner et al, investigated the impact of the pH of the acceptor fluid (pH buffer 5.5, 7.4, 8.5 and 9) on the pH of the different skin layers⁴⁷. After reaching an equilibrium of 3h with the medium, the pH of the dermis and the viable epidermis is modified, becoming close to that of the medium. A change in the pH of the skin can affect the permeation of the test substance in several ways. To maintain the viability of skin explants, some authors advise using a survival medium with a high buffering capacity to maintain a physiological pH above 5.5 for the duration of the experiment to compensate for the production of lactate by the skin (otherwise the medium must be renewed regularly)⁴⁶. Hopf et al, even recommended a pH close to 7.35²⁰.

- Donor type

The influence of the formulation or vehicle on skin penetration is evident and well documented, as certain vehicles or vehicle components help test substances to cross the SC barrier^{48,49}, modify the flux and the t_{lag} ^{13,50,51}. This is why GN156 recommends that the test preparations are similar to what humans are exposed to¹¹. But as exposure situations vary, the dataset includes innumerable donor types whose effects on the percutaneous absorption of the substance tested are different.

Measurements extracted from permeation experiments should be compared with experiments conducted on identical vehicles since K_p is a parameter that incorporates the partitioning step of the compound between the vehicle and the SC layer of the skin. A vehicle of interest in this perspective could be water but we must keep

17

in mind that water modifies some skin properties (hydration, swelling, etc.)¹³. Moreover, this raises the question of substances that are not very soluble in water, such as lipophilic substances. Below a certain solubility in water, a consensus-based vehicle other than water should be proposed.

Figure 5 illustrates donor category types included in the dataset.

- Donor pH

GN156 warns about the potential pH effects of the formulation: it can modify the ionization state of the substance tested and have deleterious effects on the skin: irritation resulting in modifying the skin's absorption parameters¹¹. However, no pH value is recommended.

The question of a donor pH is only relevant for aqueous formulations. The physiological surface pH of skin is acidic, around 5, and there is a pH gradient across the thickness of the SC⁵², with some publications indicating a gender dependence of skin pH^{47,53,54}. The pH of *in vitro* SC (frozen or fresh) is higher and can become neutral^{47,55}. The deposition solution is generally at a pH between 4 and 7, taking into account the buffering capacity of the skin⁵⁶. Caution is required as even in this range very different fluxes can be observed^{57,58}.

The buffering capacity of the skin is limited and can be overcome in case of exposure to solutions with extreme pH, as they can modify the skin barrier⁵².

Knowing that the ionized forms of a substance are much less permeable than the non-ionized forms, the pH of the donor medium will necessarily have an effect on the parameters in the case of ionizable substances at non-extreme pH, as observed for example for lignocaine flux⁵⁹.

- Occlusion

According to GN156 the choice of occlusion/non-occlusion should depend primarily on the properties of the test substance (occlusion to prevent the evaporation of volatile substances) and the exposure scenario¹¹. Generally, but not always, occlusion favors the percutaneous absorption of the test substance by increasing skin (SC) hydration and temperature, leading to a modification of the percutaneous absorption parameters⁶⁰. Bjorklund et al, showed that by decreasing the water gradient over the skin and thus increasing its hydration, the flux of 2 substances, one hydrophilic, the other lipophilic, increases drastically⁶¹. These results help to explain the effects of occlusion. Van der Merwe et al, observed the impact of occlusion on the apparent lag time⁴⁴.

- Finite-infinite dosing scenario

The K_p is calculated from the flux of the solute over the skin under steady-state conditions, i.e. in infinite dosing conditions. Indeed, steady-state is rarely reached in finite dose conditions. TGD428 advise to apply up to 10 $\mu\text{l}/\text{cm}^2$ in finite dose experiments on liquids and 100 $\mu\text{l}/\text{cm}^2$ or more in infinite dose experiments⁹. However, it is necessary to consider that these recommendations have certain limits, for example, a small volume of highly concentrated solution of a low permeated solute can behave like an infinite dose scenario⁶². Therefore, a better mathematical definition is that finite dose conditions apply when depletion of the donor occurs⁶³ with the characteristic curve shapes presented in Fig 2. Unfortunately, in practice, some researchers claim they are in infinite dose conditions but only give the deposited volume used. Several authors report a J_{ss} , K_p , and t_{lag} without mentioning whether they had previously verified that they obtained a steady-state and how they verified it.

TGD428, GD28 and GN156 do not comprehensively address methodological issues to determine the boundaries of the steady-state and the K_p in infinite dose, nor do they indicate if it possible to predict a K_p without steady-state, nor do they propose any criterion to evaluate the quality of the K_p obtained^{8,9,11}.

It is possible to extrapolate K_p from finite dose experiments but the estimated K_p are generally lower than the true values²⁰.

- Reaching steady-state

The methodology used to determine steady-state boundaries has a significant impact on the percutaneous absorption parameters as the inclusion of data collected at times before steady-state leads to underestimating both K_p and t_{lag} ⁶⁴. The time recommended for the permeation rate across a membrane to reach the steady-state value must be at least 2.7 or 3 times the lag time in order to obtain a good estimate of the permeability coefficient^{63,65,66}. Niedorf et al, proposed an automated approach based on an algorithm to define the boundaries of the steady-state⁶⁷.

- Mass-balance Recovery

At the end of the experiment, mass-balance recovery must be determined and provided (TGD 428)⁹. The GD28 and GN156 set an adequate recovery target for the test substance of 90 to 110% with a recovery of 80 to 120% tolerated for volatile and non-radiolabeled substances^{8,11}. In the case of recoveries outside this range or for non-indicated recoveries, the results obtained are questionable. Indeed, an excessively weak recovery can be due, for example, to the evaporation or adsorption of substances, particularly for lipophilic ones, on the walls of the vials or donor/acceptor compartments, or a problem of the extraction of the test substance from the skin^{20,23}. However, GD28 indicates "For infinite dose applications, a steady-state flux and a permeability coefficient (K_p) are determined. Recovery determination is not relevant because the only important end-point is the appearance of the test substance in the acceptor fluid."⁸.

This work highlights the serious need for standardization and exhaustive and comprehensive reporting of experimental conditions in skin absorption studies.

Code availability

The dataset generated (SkinPiX) is available in open source (<https://doi.org/10.57745/7FHQOY>) and the KNIME workflows used to process the data are provided there.

References

1. Brown, T. N., Armitage, J. M., Egeghy, P., Kircanski, I. & Arnot, J. A. Dermal permeation data and models for the prioritization and screening-level exposure assessment of organic chemicals. *Environ. Int.* 94, 424–435 (2016).
2. Flynn, G. L. Physicochemical determinates of skin absorption in Principles of route-to-route extrapolation for risk assessment: Proceedings. (Elsevier, 1990).
3. Williams, F. M. EDETOX. Evaluations and predictions of dermal absorption of toxic chemicals. *Int. Arch. Occup. Environ. Health* 77, 150–151 (2004).
4. Samaras, E. G., Riviere, J. E. & Ghafourian, T. The effect of formulations and experimental conditions on in vitro human skin permeation—Data from updated EDETOX database. *Int. J. Pharm.* 434, 280–291 (2012).
5. Stepanov, D., Canipa, S. & Wolber, G. HuskinDB, a database for skin permeation of xenobiotics. *Sci. Data* 7, 426 (2020).
6. Cheruvu, H. S., Liu, X., Grice, J. E. & Roberts, M. S. An updated database of human maximum skin fluxes and epidermal permeability coefficients for drugs, xenobiotics, and other solutes applied as aqueous solutions. *Data Brief* 42, 108242 (2022).
7. Roberts, M. S. et al. Topical drug delivery: History, percutaneous absorption, and product development. *Adv. Drug Deliv. Rev.* 177, 113929 (2021).
8. OECD. Test no. 28: Guidance document for the conduct of skin absorption studies. Series on testing and assessment. (OECD, 2004). doi:10.1787/9789264078796-en.
9. OECD. Test No. 428: Skin Absorption: In Vitro Method. (OECD, 2004). doi:10.1787/9789264071087-en.
10. OECD. Test no. 156 (2019): Guidance notes on dermal absorption. Series on testing and assessment (Draft Second Edition). (OECD, 2019).

11. OECD. Test no. 156 (2022): Guidance notes on dermal absorption. Series on testing and assessment (Draft Second Edition). (OECD, 2022).
12. Law, R. M., Ngo, M. A. & Maibach, H. I. Twenty Clinically Pertinent Factors/Observations for Percutaneous Absorption in Humans. *Am. J. Clin. Dermatol.* 21, 85–95 (2020).
13. Champmartin, C., Chedik, L., Marquet, F. & Cosnier, F. Occupational exposure assessment with solid substances: choosing a vehicle for in vitro percutaneous absorption experiments. *Crit. Rev. Toxicol.* 52, 294–316 (2022).
14. Zhang, A. et al. Vehicle effects on human stratum corneum absorption and skin penetration. *Toxicol. Ind. Health* 33, 416–425 (2017).
15. Berthold, M. R. et al. KNIME: The Konstanz Information Miner. in *Data Analysis, Machine Learning and Applications* (eds. Preisach, C., Burkhardt, H., Schmidt-Thieme, L. & Decker, R.) 319–326 (Springer Berlin Heidelberg, 2008). doi:10.1007/978-3-540-78246-9_38.
16. Kireeva, N. et al. Generative Topographic Mapping (GTM): Universal Tool for Data Visualization, Structure–Activity Modeling and Dataset Comparison. *Mol. Inform.* 31, 301–312 (2012).
17. Bishop, C. M., Svensén, M. & Williams, C. K. I. GTM: The Generative Topographic Mapping. *Neural Comput.* 10, 215–234 (1998).
18. Ellison, C. A. et al. Partition coefficient and diffusion coefficient determinations of 50 compounds in human intact skin, isolated skin layers and isolated stratum corneum lipids. *Toxicol. In Vitro* 69, 104990 (2020).
19. Wildman, S. A. & Crippen, G. M. Prediction of Physicochemical Parameters by Atomic Contributions. *J. Chem. Inf. Comput. Sci.* 39, 868–873 (1999).
20. Hopf, N. B. et al. Reflections on the OECD guidelines for in vitro skin absorption studies. *Regul. Toxicol. Pharmacol.* 117, 104752 (2020).
21. Sullivan, K. M. et al. Dermal absorption for pesticide health risk assessment: Harmonization of study design and data reporting for North American Regulatory submissions. *Regul. Toxicol. Pharmacol.* 90, 197–205 (2017).
22. Esposito Biondo, N., Fretes Argenta, D., Schneider Rauber, G. & Caon, T. How to define the experimental conditions of skin permeation assays for drugs presenting biopharmaceutical limitations? The experience with testosterone. *Int. J. Pharm.* 607, 120987 (2021).
23. Fabian, E., Oesch, F., Ott, K., Landsiedel, R. & van Ravenzwaay, B. A protocol to determine dermal absorption of xenobiotics through human skin in vitro. *Arch. Toxicol.* 91, 1497–1511 (2017).
24. Dąbrowska, A. K. et al. The relationship between skin function, barrier properties, and body-dependent factors. *Skin Res. Technol.* 24, 165–174 (2018).
25. Bormann, J. L. & Maibach, H. I. Effects of anatomical location on in vivo percutaneous penetration in man. *Cutan. Ocul. Toxicol.* 39, 213–222 (2020).
26. Knorr, F. et al. Follicular transport route – Research progress and future perspectives. *Eur. J. Pharm. Biopharm.* 71, 173–180 (2009).
27. Feschuk, A. M., Law, R. M. & Maibach, H. I. *Dermal Absorption and Decontamination: A Comprehensive Guide*. (Springer International Publishing, 2022). doi:10.1007/978-3-031-09222-0.
28. Poet, T. S. & McDougal, J. N. Skin absorption and human risk assessment. *Chem. Biol. Interact.* 140, 19–34 (2002).
29. Van de Sandt, J. J. M. et al. In vitro predictions of skin absorption of caffeine, testosterone, and benzoic acid: a multi-centre comparison study. *Regul. Toxicol. Pharmacol.* 39, 271–281 (2004).
30. Rothe, H. et al. Comparison of protocols measuring diffusion and partition coefficients in the stratum corneum: Diffusion and partition coefficients in the stratum corneum protocols. *J. Appl. Toxicol.* 37, 806–816 (2017).
31. Anderson, B. D., Higuchi, W. I. & Raykar, P. V. Heterogeneity Effects on Permeability–Partition Coefficient Relationships in Human Stratum Corneum. *Pharm. Res.* 05, 566–573 (1988).
32. Wilkinson, S. C. et al. Interactions of skin thickness and physicochemical properties of test compounds in percutaneous penetration studies. *Int. Arch. Occup. Environ. Health* 79, 405–413 (2006).
33. Zou, Y. & Maibach, H. I. Dermal–epidermal separation methods: research implications. *Arch. Dermatol. Res.* 310, 1–9 (2018).
34. Lau, W. M., Ng, K. W., Sakenyte, K. & Heard, C. M. Distribution of esterase activity in porcine ear skin, and the effects of freezing and heat separation. *Int. J. Pharm.* 433, 10–15 (2012).
35. Dennerlein, K. et al. Studies on percutaneous penetration of chemicals – Impact of storage conditions for excised human skin. *Toxicol. In Vitro* 27, 708–713 (2013).
36. Jacques-Jamin, C. et al. Comparison of the Skin Penetration of 3 Metabolically Stable Chemicals Using Fresh and Frozen Human Skin. *Skin Pharmacol. Physiol.* 30, 234–245 (2017).
37. Nielsen, J. B., Plasencia, I., Sørensen, J. A. & Bagatolli, L. A. Storage Conditions of Skin Affect Tissue Structure and Subsequent in vitro Percutaneous Penetration. *Skin Pharmacol. Physiol.* 24, 93–102 (2011).
38. Ng, S.-F., Rouse, J. J., Sanderson, F. D., Meidan, V. & Eccleston, G. M. Validation of a Static Franz Diffusion Cell System for In Vitro Permeation Studies. *AAPS PharmSciTech* 11, 1432–1441 (2010).
39. Nielsen, J. B., Benfeldt, E. & Holmgaard, R. Penetration through the Skin Barrier. in *Current Problems in Dermatology* (ed. Agner, T.) vol. 49 103–111 (S. Karger AG, 2016).
40. Bronaugh, R. L. & Stewart, R. F. Methods for In Vitro Percutaneous Absorption Studies IV: the Flow-Through Diffusion Cell. *J. Pharm. Sci.* 74, 64–67 (1985).
41. Clowes, H. M., Scott, R. C. & Heylings, J. R. Skin absorption: Flow-through or static diffusion cells. *Toxicol. In Vitro* 8, 827–830 (1994).
42. Kilo, S. et al. Impact of physiologically relevant temperatures on dermal absorption of active substances - an ex-vivo study in human skin. *Toxicol. In Vitro* 68, 104954 (2020).
43. Zambrana, P. N., Hou, P., Hammell, D. C., Li, T. & Stinchcomb, A. L. Understanding Formulation and Temperature Effects on Dermal Transport Kinetics by IVPT and Multiphysics Simulation. *Pharm. Res.* 39, 893–905 (2022).
44. Van der Merwe, D. et al. A Physiologically Based Pharmacokinetic Model of Organophosphate Dermal Absorption. *Toxicol. Sci.* 89, 188–204 (2006).
45. Dumont, C., Prieto, P., Asturiol, D. & Worth, A. Review of the Availability of In Vitro and In Silico Methods for Assessing Dermal Bioavailability. *Appl. Vitro Toxicol.* 1, 147–164 (2015).

46. Tarnowska, M. et al. Formulation of survival acceptor medium able to maintain the viability of skin explants over in vitro dermal experiments. *Int. J. Cosmet. Sci.* 41, 617–623 (2019).
47. Wagner, H. pH profiles in human skin: influence of two in vitro test systems for drug delivery testing. *Eur. J. Pharm. Biopharm.* 55, 57–65 (2003).
48. Abd, E., Benson, H. A. E., Mohammed, Y. H., Roberts, M. S. & Grice, J. E. Permeation Mechanism of Caffeine and Naproxen through in vitro Human Epidermis: Effect of Vehicles and Penetration Enhancers. *Skin Pharmacol. Physiol.* 32, 132–141 (2019).
49. Lane, M. E. Skin penetration enhancers. *Int. J. Pharm.* 447, 12–21 (2013).
50. Haq, A. & Michniak-Kohn, B. Effects of solvents and penetration enhancers on transdermal delivery of thymoquinone: permeability and skin deposition study. *Drug Deliv.* 25, 1943–1949 (2018).
51. Williams, A. C. & Barry, B. W. Penetration enhancers. *Adv. Drug Deliv. Rev.* 56, 603–618 (2004).
52. Parra, J. L. & Paye, M. EEMCO Guidance for the in vivo Assessment of Skin Surface pH. *Skin Pharmacol. Physiol.* 16, 188–202 (2003).
53. Luebberding, S., Krueger, N. & Kerscher, M. Skin physiology in men and women: in vivo evaluation of 300 people including TEWL, SC hydration, sebum content and skin surface pH. *Int. J. Cosmet. Sci.* 35, 477–483 (2013).
54. Man, M. Q. et al. Variation of Skin Surface pH, Sebum Content and Stratum Corneum Hydration with Age and Gender in a Large Chinese Population. *Skin Pharmacol. Physiol.* 22, 190–199 (2009).
55. Messager, S., Hann, A. C., Goddard, P. A., Dettmar, P. W. & Maillard, J.-Y. Assessment of skin viability: is it necessary to use different methodologies?: Assessment of skin viability. *Skin Res. Technol.* 9, 321–330 (2003).
56. Levin, J. & Maibach, H. Human skin buffering capacity: an overview. *Skin Res. Technol.* 14, 121–126 (2008).
57. Salocks, C. B. et al. Dermal exposure to methamphetamine hydrochloride contaminated residential surfaces. *Food Chem. Toxicol.* 50, 4436–4440 (2012).
58. Mutalik, S., Shetty, P. K., Kumar, A., Kalra, R. & Parekh, H. S. Enhancement in deposition and permeation of 5-fluorouracil through human epidermis assisted by peptide dendrimers. *Drug Deliv.* 21, 44–54 (2014).
59. Valenta, C., Siman, U., Kratzel, M. & Hadgraft, J. The dermal delivery of lignocaine: influence of ion pairing. *Int. J. Pharm.* 197, 77–85 (2000).
60. Taylor, L. J. et al. Effect of occlusion on the percutaneous penetration of linoleic acid and glycerol. *Int. J. Pharm.* 249, 157–164 (2002).
61. Björklund, S., Engblom, J., Thuresson, K. & Sparr, E. A water gradient can be used to regulate drug transport across skin. *J. Controlled Release* 143, 191–200 (2010).
62. Selzer, D., Schaefer, U. F., Lehr, C.-M. & Hansen, S. Basic Mathematics in Skin Absorption. in *Percutaneous Penetration Enhancers Drug Penetration Into/Through the Skin* (eds. Dragicevic, N. & I. Maibach, H.) 3–25 (Springer Berlin Heidelberg, 2017). doi:10.1007/978-3-662-53270-6_1.
63. Selzer, D., Abdel-Mottaleb, M. M. A., Hahn, T., Schaefer, U. F. & Neumann, D. Finite and infinite dosing: Difficulties in measurements, evaluations and predictions. *Adv. Drug Deliv. Rev.* 65, 278–294 (2013).
64. Bunge, A. L., Cleek, R. L. & Vecchia, B. E. A new method for estimating dermal absorption from chemical exposure. 3. Compared with steady-state methods for prediction and data analysis. *Pharm. Res.* 12, 972–982 (1995).
65. Selzer, D. et al. A strategy for in-silico prediction of skin absorption in man. *Eur. J. Pharm. Biopharm.* 95, 68–76 (2015).
66. Shah, J. C. Analysis of permeation data: evaluation of the lag time method. *Int. J. Pharm.* 90, 161–169 (1993).
67. Niedorf, F., Schmidt, E. & Kietzmann, M. The Automated, Accurate and Reproducible Determination of Steady-state Permeation Parameters from Percutaneous Permeation Data. *Altern. Lab. Anim.* 36, 201–213 (2008).

For modelling purposes, the published SkinPiX database was processed further by removal of unprecise values (if relation was “<”) and removal of duplicate molecules by considering median of skin permeability values.

HuskinDB

The HuskinDB (v1.01 version, August 2021) consisted of 550 data points for 253 compounds, extracted from 95 publications (1964–2012). The data processing included the removal of unreliable data, molecular standardization, manual duplicate processing, and application of the exclusion criteria defined during the curation of the SkinPiX database. The decision of removal of unreliable data and the exclusion criteria were defined by skin permeability experts from the INRS. The exclusion is based on parameters, namely skin source site, used skin layer, skin preparation method, donor type, acceptor type, and cell type. After molecular standardization (described in the section “Molecular standardization” of the “QSAR/QSPR modelling methodology”), duplicate molecules were manually examined and irrelevant duplicates were removed by the suggestion of our collaborators from INRS. For the remaining duplicates, the median value was taken. The final step was application of the exclusion criteria formed during the processing of the SkinPiX database, which were also related to such parameters like, skin preparation, used skin layer, etc. The resulting curated HuskinDB contained 128 compounds.

During the course of modelling, 5 potential outliers were identified and discarded. The final training set was composed of 123 compounds (Figure 13).

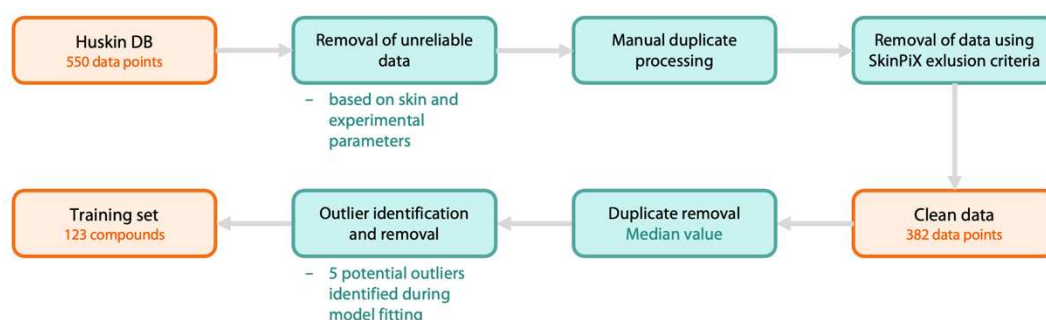


Figure 13. Data processing of the HuskinDB.

Merged training set preparation

The merged training set is a combination of the HuskinDB and SkinPiX, which was created with an intention of building a QSPR model with a larger applicability domain. The data processing involved duplicate processing. Duplicates were merged using the median value for their label. This yielded 203 compounds in the merged dataset. During modelling stage, 8 outliers were identified and removed, leaving 195 compounds in the dataset.

5.3.3 Methods

Modelling workflow

The model training and validation was performed using 5-fold external cross-validation. SVM models were trained on ISIDA fragment descriptors of various topologies (sequences, atom-centered fragments, triplets) and lengths (from 2 to 3 atoms). The hyperparameters were optimized using hill-climbing method. All modelling steps were performed using KNIME workflows.

Outlier identification and removal

The outlier detection and removal were performed due to low performance of models during 5-fold cross-validation. It involved selection of the best performing descriptor set (decided based on 5-fold cross-validation) and fitting of the model to the whole training set. The compounds with the difference between predicted and experimental value greater than or equal to 1 log were considered as outliers. Once they are removed, the 5-fold cross-validation is performed again and the presence of outliers was verified. This procedure was performed until no outlier was detected.

Generative topographic mapping

GTM was used to visualize the chemical space coverage by the training and test sets. This was achieved by training GTM model on a ISIDA fragment descriptor set, namely, atom-centered fragments of fixed length ranging from 2 to 3 atoms radius.

5.3.4 Results and discussion

The 5-fold cross-validation performance of the model built on HuskinDB is provided in Figure 14. The removal of 5 identified outliers resulted in improvement of performance. The observed $R^2_{5-CV} = 0.53$ is similar to the performance of the model trained on HuskinDB ($R^2_{test} = 0.5$) published by Waters and Quah⁸⁹. The consensus model built on HuskinDB consisted of 7 individual models with internal validation R^2 ranging from 0.64 to 0.78.

The consensus model showed poor performance when applied to the external test set, SkinPiX data ($R^2_{test} = -0.21$, $RMSE_{test} = 1.32$) (Figure 15). In order to understand the reason of poor prediction, the chemical space coverage of the HuskinDB model training set and the SkinPiX test set was analyzed using GTM class landscape (Figure 16). The class landscape clearly indicates the zones populated by one of the datasets (blue color for the training set; red color for the test set), showing the limited applicability domain of the HuskinDB model.

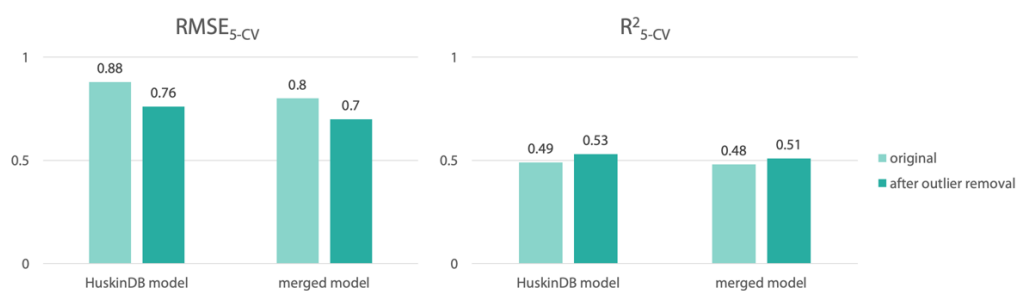


Figure 14. 5-fold cross-validation performance of "HuskinDB model" and "merged model".

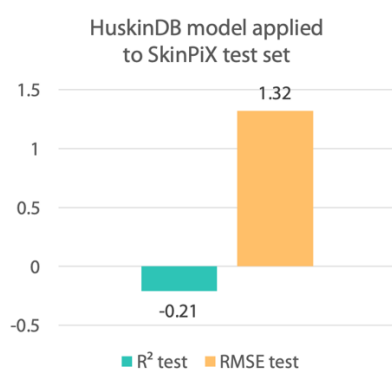


Figure 15. Performance of "HuskinDB model" on SkinPiX test set.

In order to expand the applicability domain, a new model was trained on the merged dataset composed of both the SkinPiX and the HuskinDB datasets. The 5-fold cross-validation performance increased after the removal of 8 outliers. The RMSE of the merged model (0.7) aligns well with the mean standard deviation of duplicate molecules (0.68). A better understanding of this case requires a more thorough investigation of the merged dataset, specifically, during the data blending stage.

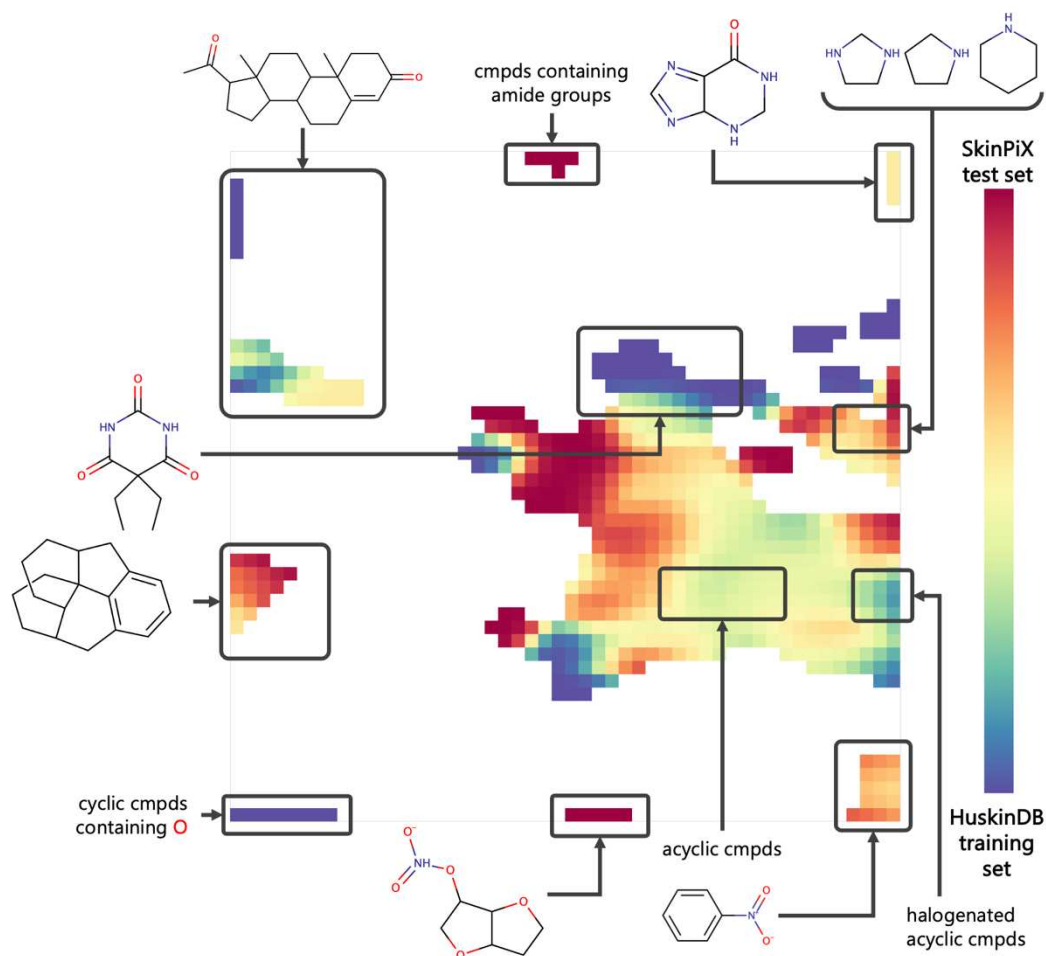


Figure 16. GTM class landscape showing the distribution of the HuskinDB training dataset (blue) and the SkinPiX test set (red). Yellow and green areas are populated with compounds of both datasets. The structural motifs found in certain zones are displayed.

5.3.5 Summary

Overall, this study resulted in the compilation of a new skin permeability database, called SkinPiX¹³. This database is a collection of skin permeability and assisting metadata (experimental set-ups, etc.) extracted from articles published between 2012 and 2021. Public availability of the new database will facilitate the research conducted in the skin permeability domain as well as enhance the development of novel QSPR models that cover a broader chemical space, consequently improving the accuracy of predictions. The developed model is freely available on the web service of the Laboratory of Chemoinformatics⁶⁹ (“Skin permeability - Classification” model in the “PhysProp” section).

The modelling results of this work can be extended by a more careful study of SkinPiX and HuskinDB databases, in order to define a better ground for merging the relevant data. Identification of the criteria for merging the datasets, will improve the homogeneity of the

merged training set, hence, it will enhance the predictive performance of the model. The predictions of SkinPiX data using HuskinDB based models were poor. The prediction of HuskinDB data using the SkinPiX data remains to be done. Additionally, a comprehensive analysis of the outliers is still missing. Yet, the trend is that the current datasets are so small that the models are quite unstable and require a very stringent applicability domain to be used. Additionally, detecting outliers and inconsistencies in the data is likely to be strategic in improving the quality of skin permeability QSPR models.

5.4 Conclusion

The findings of this chapter underscore the significant advancements in skin-related safety properties assessment. The comparative study has convincingly demonstrated the superior performance of the bone marrow-derived dendritic cell (BMDC) assay in evaluating skin sensitization compared to other existing tests. The BMDC assay and the developed QSAR model provide experts with an efficient approach to prioritize compound analysis, reducing time, resources, and material.

Moreover, the meticulous compilation of the new skin permeability database, SkinPiX, adds valuable insights to the field. By encompassing a comprehensive collection of skin permeability data and essential metadata from published articles between 2012 and 2021, SkinPiX offers a crucial resource for the development of novel QSPR models, ensuring a broader coverage of chemical space and improved accuracy in predicting skin permeability.

All skin permeability and sensitization data and models are publicly available^{13,69}. The results of both skin permeability and sensitization projects are the product of a successful collaboration with the Institut National de Recherche et de Sécurité (INRS) in Nancy.

Chapter 6

ACE2 selective inhibition

6.1 Introduction

Angiotensin-converting enzyme 2 (ACE2) is an enzyme found in various cell types, including pulmonary, cardiac, and renal cells. It plays a vital role in the renin-angiotensin-aldosterone system (RAAS), which regulates blood pressure and fluid balance in the body.¹⁴ In addition, ACE2 acts as the primary receptor for severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), enabling the virus to enter and infect human cells.¹⁵ Understanding the molecular mechanisms of ACE2, including its role in SARS-CoV-2 infection, is therefore essential for monitoring its influence on different biological tissues and organs. Chemical probes, as small molecules with selective binding capabilities, provide a valuable tool for investigating mechanistic and phenotypic questions about ACE2 in various biological studies, allowing researchers to gain insights into the target's function and potential therapeutic implications.⁹⁰⁻⁹² Development of ACE2-targeting probes with selectivity towards ACE2 over closely related targets such as ACE and neprilysin (NEP), is critical to differentiate the effects caused by ACE2 downregulation from the effects of interacting with these other targets, particularly in the context of blood pressure regulation.

The aim of this project is to identify new potential ACE2 selective inhibitors to be suggested for the secondary screening campaign. The advancement builds upon previous efforts to investigate ACE2 inhibition and develop selective chemical probes.⁹³ Virtual screening techniques were employed to explore Enamine's vast compound collection of 2.6 million compounds and a set of 4080 *in silico* designed compounds⁹⁴ to generate a list of promising ACE2 binders. Virtual screening methods included docking, structure-based pharmacophore and QSAR modelling. New QSAR classification models predicting the inhibition of ACE2, ACE, and NEP enzymes were developed and uploaded to the Predictor web service⁶⁹ of the Laboratory of Chemoinformatics (“ACE2 inhibition -

Classification”, “ACE inhibition - Classification”, “NEP inhibition - Classification” models in the “Activity” section). The overall workflow of the project and the role of QSAR models in it is presented in Figure 17.

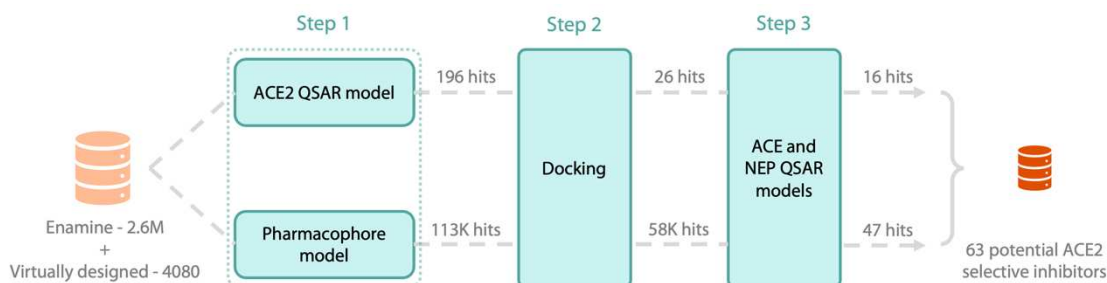


Figure 17. Overview of virtual screening steps to identify selective ACE2 inhibitors.

6.1.1 State-of-the-art

Literature review revealed several *in silico* studies in the domain of ACE2 inhibition, including approaches like molecular docking⁹⁵, QSAR model training on docking scores⁹⁶, 3D QSAR pharmacophore modelling⁹⁷. The most similar work, however, was conducted by Hochuli et al.⁹⁸, where both computational and experimental approaches were combined. Their goal was to identify allosteric ACE2 binders that would have the potential to serve as a novel class of antiviral agents for the treatment of COVID-19. First, they performed experimental screening to test compounds for ACE2 binding and enzymatic inhibition. QSAR models were then used to predict and prioritize compounds, followed by ligand-based pharmacophore modeling to select additional candidates. Subsequent experimental validation confirmed that 5 compounds exhibited strong ACE2 binding to an allosteric site, minimal enzymatic inhibition, and significant inhibition of SARS-CoV-2 replication in human cells.

This project is a direct continuation of the previous study⁹³ conducted in collaboration with the Institute of Organic Chemistry and Enamine (Kyiv, Ukraine). The workflow of the project involved primary screening of molecules identified by virtual screening methods. Firstly, molecular docking, structure-, and ligand-based pharmacophore modeling were employed on the Enamine in-stock compound collection to identify potential selective ACE2 inhibitors. Subsequently, QSAR models predicting ACE and NEP inhibition were constructed and applied to the virtual hits obtained from the previous approaches. The list of 577 virtual hits obtained from the Laboratory of Chemoinformatics (Strasbourg, France), the computational chemistry team of Enamine (Kyiv, Ukraine) and Chemspace LLC (Kyiv, Ukraine) was submitted for experimental validation. Although none of the compounds showed activity in the nanomolar region, two of them possessed optimal parameters for penetrating blood-brain barrier. Those ligands also displayed novel ACE2-binding chemotypes and have the potential to become more efficient with further structural optimization. ACE and NEP activities were not experimentally assessed.

The results of the primary screening were used in this project, to update and improve the performance of models, consequently leading for more potent virtual hits. The newly identified virtual hits are submitted for the second experimental validation. The project was conducted together with my colleague Ms. Farah Asgarkhanova, Ph.D. student.

6.2 Data

The dataset for ACE2 was compiled from experimentally validated molecules provided by Enamine (552 data points), ChEMBL (release 30; Target ID = 3736) (100 data points), and PubChem (71 data points). The source of data for ACE (Target ID = ChEMBL1808; 1108 data points) and NEP (Target ID = ChEMBL1944; 694 data points) was ChEMBL (release 30). The origins of enzymes in all cases were mammals. The molecules reported either IC₅₀, K_i or percentage inhibition.

The molecular standardization was performed using the ChemAxon Standardizer⁵⁴. The standardization protocol included dearomatization, dealkalization, removal of salts and mixtures, neutralization, generation of the major tautomer, aromatization. After the standardization duplicate molecules were removed. The values and units were converted from K_i and IC₅₀ to pK_i and pIC₅₀, respectively.

A compound was classified as ACE2 inactive if its percentage inhibition was less than 25% or if pK_i or pIC₅₀ was less than 8, and classified as ACE2 active otherwise. For ACE and NEP, classification threshold was less strict, with 60% inhibition and pK_i or pIC₅₀ = 6. The final training set sizes are given in Table 7.

Table 7. Training set sizes of ACE2, ACE and NEP models.

Enzyme	Training set size	Number of actives	Number of inactives
ACE2	668	37	631
ACE	591	304	287
NEP	464	301	163

The datasets which were screened were composed of 2.6 million Enamine compounds and 4080 compounds generated by the Synt-On tool⁹⁴. The virtually designed dataset was generated based on 37 experimentally confirmed active compounds identified during the first screening campaign.

6.3 Methods

Methods used for identification of ACE2 selective binders were applied in 3 steps (Figure 17):

1. Independent application of ACE2 QSAR and pharmacophore modeling approaches.
2. Molecular docking of virtual hits predicted by QSAR and pharmacophore models from Step 1.
3. Filtering virtual hits obtained from Step 2, by applying ACE and NEP QSAR models to find ACE2-selective virtual hits.

6.3.1 QSAR modeling

QSAR models were trained on ISIDA fragment descriptors using SVM machine learning method and hyperparameters optimized by genetic algorithm. Validation of models was performed using 5-fold cross-validation technique. Based on the cross-validation performance, top 7 best performing models were selected for consensus model (Table 8). ACE2, ACE and NEP models were used as filters to sieve irrelevant hits throughout virtual screening process.

Table 8. Performance of ACE2, ACE and NEP consensus QSAR models, and their constituting models. Fragmentation types: I – sequence; II – atom-centered; III – triplet; A – atom; B – bond; R – fragment of fixed length; P – “Atom Pairs” option; AP – “Do All Ways” option. BA_{5-cv} – balanced accuracy on 5-fold cross-validation.

Performance of individual models					
ACE2		ACE		NEP	
Descriptor set	BA _{5-cv}	Descriptor set	BA _{5-cv}	Descriptor set	BA _{5-cv}
IIA(2-6)_R	0.97	IIA(2-4)	0.845	IIA(2-7)_R	0.81
IIAB(2-4)_R	0.97	IIA(3-5)_R	0.84	IIAB(3-6)_R	0.81
IIAB(2-7)_R_P	0.97	IIAB(3-5)_R	0.835	IAB(2-6)	0.79
IIA(3-4)	0.97	IIAB(3-5)	0.835	III(3-6)	0.79
IIAB(3-7)_R	0.965	IIA(2-5)	0.83	IIA(3-4)_P	0.78
IIAB(3-4)_P	0.96	IIAB(2-6)_R_P	0.83	IA(2-5)_AP	0.78

Performance of consensus models (BA _{5-cv})		
ACE2	ACE	NEP
0.97	0.83	0.79

6.3.2 Rule-based algorithm

A rule-based algorithm derives patterns from data using a set of *if-then* logical statements. In the context of this study, we employed JRip rule-based classification method, which is an extension of RIPPER (Repeated Incremental Pruning to Produce Error Reduction) algorithm⁹⁹. It constructs a set of *if-then* rules to predict class labels for data instances. Beginning with a single rule that predicts the majority class, JRip iteratively adds and prunes rules, using a heuristic to select the most informative attribute to augment a rule and the least impacting attribute to prune a rule. This process continues until a sufficiently accurate and interpretable rule set is established, providing a transparent framework for classification tasks. In our case, we used JRip method to find substructural motifs that are responsible for ACE2 inhibition. The method was implemented in Weka software¹⁰⁰.

6.3.3 Pharmacophore modeling

Pharmacophore modeling involves the identification and characterization of essential chemical features (hydrogen bond donors/acceptors, aromatic rings, etc.) and spatial arrangements within a molecule that are critical for binding to a target receptor or enzyme. Structure-based pharmacophore model integrates structural information of the target receptor to find pharmacophoric features for a specific binding site. Once a set of pharmacophoric features is identified, it can be used as a query to screen libraries of compounds. Compounds that optimally fit these features are considered virtual hits. In this study, the ACE2 protein (PDB ID: “1R4L”¹⁰¹) was obtained from the Protein Data Bank (PDB). LigandScout¹⁰² (v. 4.4.8) was used to generate and apply pharmacophore models. The pharmacophore model was used as a virtual screening step.

6.3.4 Molecular docking

Molecular docking approach is used to assess the binding poses and energies between ligands and target proteins. A docking score is used to model the strength of supramolecular interactions. The PLANTS (Protein-Ligand ANT System)¹⁰³ employed in this study, utilizes ant colony optimization algorithm to explore the vast conformational space of ligand-receptor complexes. This method effectively balances exploration and exploitation to identify energetically favorable binding conformations. A key component of PLANTS is the integration of the CHEMPLP scoring function¹⁰⁴, which evaluates molecular interactions, electrostatics, van der Waals forces, and solvation effects to provide accurate estimates of binding affinities. The docking was based on the 3Å resolution X-ray structure of the ACE2 binding site (1R4L) and the co-crystallized ligand (MLN-4760). The virtual hits obtained from ACE2 QSAR and pharmacophore models application step were docked onto the ACE2 binding site.

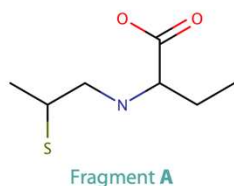
6.4 Results and discussion

As mentioned in 6.3 Methods section, the workflow of the project can be summarized in 3 steps (Figure 17). The results are organized in the same manner: (1) ACE2 QSAR and pharmacophore modeling; (2) molecular docking; (3) ACE and NEP QSAR modeling.

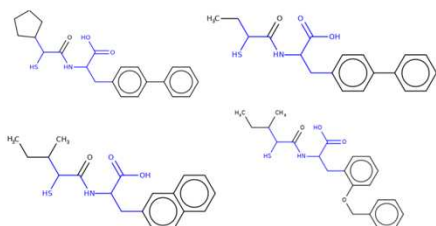
6.4.1 ACE2 QSAR and pharmacophore modeling

Prior to application of ACE2 QSAR model, the high performance of the model ($BA_{5-CV} = 0.97$) was investigated by using JRip rule-based algorithm on the ACE2 dataset. The results revealed 2 rules presented in Figure 18. The derivation of such simple rules was possible due to representation of inhibitors class by a homogeneous series of compounds. The 2 rules in Figure 18 cover 35 out of the 37 actives and 630 out of 631 inactives in the training set. These rules identified scaffolds specific to the active set of compounds. They explain the high performances ACE2 QSAR models, and anticipate potential applicability domain problems.

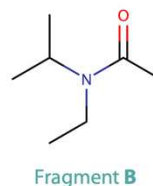
Rule 1 → Fragment A ≥ 1 and # of carbon atoms ≤ 22 ⇒ Inhibitor



Example molecules



Rule 2 → Fragment B ≥ 1 ⇒ Inhibitor



Example molecules

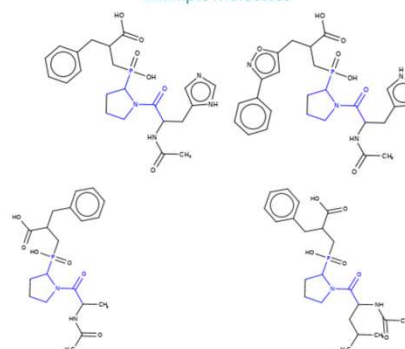


Figure 18. Classification rules derived by the JRip rule-based algorithm for ACE2 inhibition.

The application of ACE2 QSAR model to Enamine compound collection and *in silico* designed molecules resulted in 196 molecules classified as active (188 from Enamine; 8 from the generated dataset). The low number of hits is due to the bias of the model towards inactive compounds, which is dictated by the class distribution in its training set.

The LigandScout software generated structure-based pharmacophore including 10 features: 2 hydrophobic regions (H), 5 hydrogen bond donors (HD), 2 hydrogen bond acceptors (HA), 1 halogen donor (XD). The pharmacophore was optimized regarding to the “actives/hits” ratio. The final pharmacophore model included 6 features: 3 HD, 1 HA, 2H (Figure 19).

The application of the pharmacophore model to Enamine compound collection and *in silico* designed molecules resulted in 113413 molecules predicted as active (113258 from Enamine; 155 from the generated dataset).

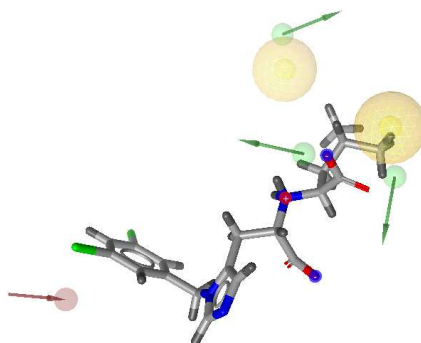


Figure 19. Structure-based pharmacophore model. Red spheres – H-bond acceptor; green spheres – H-bond donor; yellow spheres – hydrophobic regions.

6.4.2 Molecular docking

Firstly, the co-crystallized ligand and the ACE2 training set were docked. The docking score of the co-crystallized ligand was -103.48. The docking score threshold was optimized to find the optimal ratio of true actives among those having a docking score lower than the threshold. The value -85 was selected corresponding to an enrichment factor equal to 1.45. The ROC AUC score of ACE2 training compounds sorted by docking score was of 0.75.

The application of the molecular docking method to the results from the previous step (196 hits from the ACE2 QSAR model; 113413 hits from the pharmacophore model) resulted in 58185 hits.

6.4.3 ACE and NEP QSAR modeling

The final step involved application of ACE and NEP QSAR models in order to identify compounds that do not bind neither to ACE nor to NEP. In total, 63 hits were selected out of 58185 compounds that were identified at the previous virtual screening step.

6.5 Conclusion

This study resulted in successful identification of 63 potential ACE2 inhibitors from Enamine's compound collection and *in silico* designed compounds. Experimental validation of these potential ACE2 inhibitors is currently underway. The confirmed ACE2 inhibitors will set a base for the development of chemical probes that will offer valuable insights into the effects related to inhibition of ACE2.

The training set of new ACE2 model has been expanded by incorporating experimentally confirmed results from the first screening campaign of the Enamine as well as new data from ChEMBL and PubChem. This enlarged training set enables a wider applicability domain and improves the performance of the models. The freely available ACE2, ACE and NEP QSAR models can facilitate other related screening projects by virtually filtering irrelevant and prioritizing promising hits.

Overall, this work demonstrated an example of a good synergy between virtual screening and experimental validation, which led to identification of potential selective chemical probes that will shed light on potential side effects associated with ACE2 perturbation.

Chapter 7

Conclusion and perspectives

The achievements of this thesis can be summarized in two main points: development of 8 publicly available predictive models (Table 10) and user-friendly automatized KNIME workflows for building QSAR/QSPR models. Both the developed predictive models and the modelling workflows provide vast number of possibilities on their applications.

The models have demonstrated their applicability in various steps of screening by using predictive models to prioritize testing of certain compounds; to evaluate stock solution integrity; to assess the quality of experimental data; to annotate chemical libraries, prioritize compounds and identify suspicious hits and non-hits for screening campaign. Predictions made by these models are labelled based on the confidence of prediction, providing users with additional information for decision-making. The summary of the models is given in Table 6. All of the models developed in the course of this thesis are publicly available at the web service of the Laboratory of Chemoinformatics⁶⁹. The means of accessing the models are presented in Table 11 and in Figure 20.

The KNIME workflows provide necessary tools to develop QSAR/QSPR models and covers all aspects of the modelling pipeline: molecular standardization, molecular descriptor calculation, model training and validation, preparation and deployment of consensus model. The workflow manual documents and the visual programming aspects of the KNIME Analytics Platform allow users with limited coding knowledge to easily utilize these workflows and build in-house models based on their proprietary or public datasets. The KNIME workflows are available upon request to the Laboratory of Chemoinformatics: https://infochim.chimie.unistra.fr/?page_id=11.

All data are published following the FAIR principles in the repository France Data Gouv (<https://entrepot.recherche.data.gouv.fr/dataverse/CI>) or in the supplementary

Conclusion and perspectives

materials of publications (Table 9). ACE2, ACE and NEP datasets will be published after the completion of experimental validation.

Table 9. The list of published data and links to access them.

Property / Activity	Link
Solubility in DMSO	https://doi.org/10.3390/molecules26133950
Aqueous kinetic solubility	https://doi.org/10.57745/ZWS0WC
Aqueous thermodynamic solubility	https://doi.org/10.57745/CZVZIA
Skin permeability	https://doi.org/10.57745/7FHOOY
Skin sensitization	https://doi.org/10.57745/PPAMKY
ACE2 inhibition	(to be published after experimental validation)
ACE inhibition	(to be published after experimental validation)
NEP inhibition	(to be published after experimental validation)

Table 10. The list of developed QSAR/QSPR models, their training set sizes and predictive performance values. The size of the test set is indicated in brackets. BA – balanced accuracy; CV – cross-validation; RMSE – root mean-squared error.

Property / Activity	Training set size	Validation method	Performance (BA)
Solubility in DMSO	788	5-fold CV	0.78
Aqueous kinetic solubility	56132	Test set (17666)	0.84
Skin sensitization	117	5-fold CV	0.82
ACE2 inhibition	668	5-fold CV	0.97
ACE inhibition	591	5-fold CV	0.83
NEP inhibition	464	5-fold CV	0.79
Property / Activity	Training set size	Validation method	Performance (RMSE)
Skin permeability	195	5-fold CV	0.7
Aqueous thermodynamic solubility*	42159	Test set (5728)	0.59

* The QSPR model was developed by my colleague Mr. Pierre Llompart, Ph.D. student.

Table 11. The list of developed models and how to access them. All models (except thermodynamic aqueous solubility) are available on the <https://chematlas.chimie.unistra.fr/cgi-bin/predictor2.cgi> web page. The models can be accessed by first selecting "General kind of property" and then "Property to model".

Property / Activity	"General kind of property"	"Property to model"
Solubility in DMSO	PhysProp	Solubility_DMSO_2CIs
Aqueous kinetic solubility	PhysProp	Kinetic_solubility_2CIs
Skin permeability	PhysProp	Skin_permeability_Reg
Skin sensitization	Activity	Skin_sensitization_BMDC_2CIs
ACE2 inhibition	Activity	ACE2_2CIs
ACE inhibition	Activity	ACE_2CIs
NEP inhibition	Activity	NEP_2CIs
Aqueous thermodynamic solubility*	-	-

* The QSPR model was developed by my colleague Mr. Pierre Llompart, Ph.D. student. The model is available on a separate web page: https://chematlas.chimie.unistra.fr/WebTools/predictor_solubility.php

Predictor

Selected model : Solubility_DMSO_2CIs

Please note that only the first 100 molecules have been sent to the Predictor.

All calculations processed!

Thank you for your patience.

[Explore results with the scrollbar](#)

Molecule Id	Molecule Name	Predicted value	Applied models	Prediction confidence	2D structure	Comments
1		Soluble	45/45			Optimal prediction confidence : AD satisfied for 100% of individual models and repartition of classes= 95.56% .

[Back to Main Menu](#)

[Download results](#)

Figure 20. Screenshots showing example of request for ISIDA Predictor web service. Image A shows the ISIDA Predictor configuration page, where a user can select “Activity”/“PhysProp” general kind of property and then choose the model of interest. Image B illustrates an output of ISIDA Predictor. Color code of prediction confidence is as follows: green – optimal; blue – good; orange – average; red – unreliable.

7.1 Perspectives

This work can be further extended with the generation of new models covering other screening relevant properties, such as cytotoxicity¹⁰⁵ and permeability¹⁰⁶. Assessing of cytotoxicity is important as it is often conducted prior to many bioassays to check cell viability. Once cell viability is confirmed, the compound can be evaluated for its permeability properties.

The list of skin-related QSAR models can be also augmented by modeling other parameters, such as maximum skin flux, skin penetration enhancement, diffusion coefficients in different layers of skin.¹¹ The cooperation with INRS provides many other relevant targets in order to investigate danger and estimate the risks, for instance neurotoxicity and blood-brain barrier permeability. In a simplistic view, the neurotoxicity estimates a danger, but the risk evaluation requires to estimate the permeability.

The developed models can be used for profiling of screening libraries, such as Chimiothèque Nationale³⁸. In a first step, the current chemical library shall be annotated using these models, then these models shall be given to ChemBioFrance to maintain the annotations. The initial annotations being provided for efficiency and as reference to compare to when the models will be installed on the servers of the Chimiothèque Nationale in Montpellier. Property profiles will add value to the Chimiothèque Nationale.

In terms of technological advancement, active learning approach could be applied to the screening context.¹⁰⁷ The idea is to iteratively select the most informative compounds for labeling and then adding the new experimental data to the training set of the model to increase its predictive performance. This approach maximizes the use of resources and accelerates the screening process, making it particularly valuable when dealing with large chemical libraries or limited experimental capacity. However, the implementation of active learning strategies requires a close collaboration with screening platforms, as the direct access to the analytical instruments would be necessary.

Another possible technical improvement could be made by developing KNIME workflows to perform clustering for the analysis and selection of candidates for the next screening campaigns. This would be useful in the case of a cascade of sequential screening campaigns. However, the realization of this idea would again require direct access to screening platform capabilities. Some instances of such integration have already been reported.^{108–110}

Finally, the developed models in combination with other *in silico*, *in vitro*, or *in chemico* assays could be suggested as a replacement to animal testing. Such approach is already practiced in skin sensitization domain.^{111,112} In Europe, the approval of replacing an animal test with the alternative method is regulated by European Union Joint Research Centre for Alternatives to Animal Testing (EURL ECVAM).^{85 113,114} However, it is debatable whether the results of alternative *in silico*, *in vitro*, or *in chemico* methods, either alone or in combination, can ever truly represent human data. For such replacements to occur, there would need to be full transitivity between predicted / experimental data and animal / human data, which seems presently out of reach due to the complex nature of biological systems. Yet, the frontier is definitely moving in this direction.

References

- (1) Bajorath, J. Integration of Virtual and High-Throughput Screening. *Nat Rev Drug Discov* **2002**, *1* (11), 882–894. <https://doi.org/10.1038/nrd941>.
- (2) Berthold, M. R.; Cebron, N.; Dill, F.; Gabriel, T. R.; Kötter, T.; Meinel, T.; Ohl, P.; Sieb, C.; Thiel, K.; Wiswedel, B. KNIME: The Konstanz Information Miner. In *Studies in Classification, Data Analysis, and Knowledge Organization (GfKL 2007)*; Springer, **2007**.
- (3) Leach, A. R.; Hann, M. M.; Burrows, J. N.; Griffen, E. J. Fragment Screening: An Introduction. *Mol. Bio.Syst.* **2006**, *2* (9), 429. <https://doi.org/10.1039/b610069b>.
- (4) Tetko, I. V.; Novotarskyi, S.; Sushko, I.; Ivanov, V.; Petrenko, A. E.; Dieden, R.; Lebon, F.; Mathieu, B. Development of Dimethyl Sulfoxide Solubility Models Using 163 000 Molecules: Using a Domain Applicability Metric to Select More Reliable Predictions. *J. Chem. Inf. Model.* **2013**, *53* (8), 1990–2000. <https://doi.org/10.1021/ci400213d>.
- (5) Kerns, E. H.; Di, L. Solubility. In *Drug-like Properties: Concepts, Structure Design and Methods*; Elsevier, **2008**; pp 56–85. <https://doi.org/10.1016/B978-012369520-8.50008-5>.
- (6) Alsenz, J.; Kansy, M. High Throughput Solubility Measurement in Drug Discovery and Development. *Advanced Drug Delivery Reviews* **2007**, *59* (7), 546–567. <https://doi.org/10.1016/j.addr.2007.05.007>.
- (7) Ezendam, J.; Braakhuis, H. M.; Vandebriel, R. J. State of the Art in Non-Animal Approaches for Skin Sensitization Testing: From Individual Test Methods towards Testing Strategies. *Arch Toxicol* **2016**, *90* (12), 2861–2883. <https://doi.org/10.1007/s00204-016-1842-4>.
- (8) OECD. The Adverse Outcome Pathway for Skin Sensitisation Initiated by Covalent Binding to Proteins, **2014**. <https://www.oecd-ilibrary.org/content/publication/9789264221444-en>.
- (9) Battais, F.; Huppert, C.; Langonné, I.; Muller, S.; Sponne, I. In *Vitro* Detection of Chemical Allergens: An Optimized Assay Using Mouse Bone Marrow-Derived Dendritic Cells: IN VITRO DETECTION OF CHEMICAL ALLERGENS. *Contact Dermatitis* **2017**, *77* (5), 311–322. <https://doi.org/10.1111/cod.12829>.
- (10) Ng, K. W.; Lau, W. M. Skin Deep: The Basics of Human Skin Structure and Drug Penetration. In *Percutaneous Penetration Enhancers Chemical Methods in Penetration Enhancement*; Dragicevic, N., Maibach, H. I., Eds.; Springer Berlin Heidelberg: Berlin, Heidelberg, **2015**; pp 3–11. https://doi.org/10.1007/978-3-662-45013-0_1.
- (11) Tsakovska, I.; Pajeva, I.; Al Sharif, M.; Alov, P.; Fioravanzo, E.; Kovarich, S.; Worth, A. P.; Richarz, A.-N.; Yang, C.; Mostrag-Szlichtyng, A.; Cronin, M. T. D. Quantitative Structure-Skin Permeability Relationships. *Toxicology* **2017**, *387*, 27–42. <https://doi.org/10.1016/j.tox.2017.06.008>.
- (12) Stepanov, D.; Canipa, S.; Wolber, G. HuskinDB, a Database for Skin Permeation of Xenobiotics. *Sci Data* **2020**, *7* (1), 426. <https://doi.org/10.1038/s41597-020-00764-z>.
- (13) Chedik, L.; Baybekov, S.; Cosnier, F.; Marcou, G.; Varnek, A.; Champmartin, C. Données de Réplication Pour: SkinPiX (Skin Permeation of Identified Xenobiotics): An Update of Skin Permeability Data Based on A Systematic Review of Recent Research, **2023**. <https://doi.org/10.57745/7FHQOY>.
- (14) Zisman, L. S. ACE and ACE2: A Tale of Two Enzymes. *European Heart Journal* **2005**, *26* (4), 322–324. <https://doi.org/10.1093/eurheartj/ehi043>.
- (15) Clausen, T. M.; Sandoval, D. R.; Spliid, C. B.; Pihl, J.; Perrett, H. R.; Painter, C. D.; Narayanan, A.; Majowicz, S. A.; Kwong, E. M.; McVicar, R. N.; Thacker, B. E.; Glass, C. A.; Yang, Z.; Torres, J. L.; Golden, G. J.; Bartels, P. L.; Porell, R. N.; Garretson, A. F.; Laubach, L.; Feldman, J.; Yin, X.; Pu, Y.; Hauser, B. M.; Caradonna, T. M.; Kellman, B. P.; Martino, C.; Gordts, P. L. S. M.; Chanda, S. K.; Schmidt, A. G.; Godula, K.; Leibel, S. L.; Jose, J.; Corbett, K. D.; Ward, A. B.; Carlin, A. F.; Esko, J. D. SARS-CoV-2 Infection Depends on Cellular Heparan Sulfate and ACE2. *Cell* **2020**, *183* (4), 1043–1057.e15. <https://doi.org/10.1016/j.cell.2020.09.033>.
- (16) Wermuth, C. G. Chapter 6 - Strategies in the Search for New Lead Compounds or Original Working Hypotheses. In *The Practice of Medicinal Chemistry (Third Edition)*; Wermuth, C. G., Ed.; Academic Press: New York, **2008**; pp 123–143. <https://doi.org/10.1016/B978-0-12-374194-3.00006-8>.
- (17) Sergienko, E. A. Basics of HTS Assay Design and Optimization. In *Chemical Genomics*; Fu, H., Ed.; Cambridge University Press: Cambridge, **2012**; pp 159–172. <https://doi.org/10.1017/CBO9781139021500.017>.
- (18) Wildey, M. J.; Haunso, A.; Tudor, M.; Webb, M.; Connick, J. H. Chapter Five - High-Throughput Screening. In *Annual Reports in Medicinal Chemistry*; Goodnow, R. A., Ed.; Platform Technologies in Drug Discovery and Validation; Academic Press, **2017**; Vol. 50, pp 149–195. <https://doi.org/10.1016/bs.armc.2017.08.004>.

- (19) Kerns, E. H.; Di, L. Physicochemical Profiling: Overview of the Screens. *Drug Discovery Today: Technologies* **2004**, *1* (4), 343–348. <https://doi.org/10.1016/j.ddtec.2004.08.011>.
- (20) Di, L.; Kerns, E. H. Chapter 25 - Solubility Methods. In *Drug-Like Properties (Second Edition)*; Di, L., Kerns, E. H., Eds.; Academic Press: Boston, **2016**; pp 313–324. <https://doi.org/10.1016/B978-0-12-801076-1.00025-3>.
- (21) Murray, C. W.; Rees, D. C. The Rise of Fragment-Based Drug Discovery. *Nature Chem* **2009**, *1* (3), 187–192. <https://doi.org/10.1038/nchem.217>.
- (22) Kirsch, P.; Hartman, A. M.; Hirsch, A. K. H.; Empting, M. Concepts and Core Principles of Fragment-Based Drug Design. *Molecules* **2019**, *24* (23), 4309. <https://doi.org/10.3390/molecules24234309>.
- (23) Jhoti, H.; Williams, G.; Rees, D. C.; Murray, C. W. The “rule of Three” for Fragment-Based Drug Discovery: Where Are We Now? *Nat Rev Drug Discov* **2013**, *12* (8), 644–644. <https://doi.org/10.1038/nrd3926-c1>.
- (24) Lau, W. F.; Withka, J. M.; Hepworth, D.; Magee, T. V.; Du, Y. J.; Bakken, G. A.; Miller, M. D.; Hendsch, Z. S.; Thanabal, V.; Kolodziej, S. A.; Xing, L.; Hu, Q.; Narasimhan, L. S.; Love, R.; Charlton, M. E.; Hughes, S.; van Hoorn, W. P.; Mills, J. E. Design of a Multi-Purpose Fragment Screening Library Using Molecular Complexity and Orthogonal Diversity Metrics. *J Comput Aided Mol Des* **2011**, *25* (7), 621–636. <https://doi.org/10.1007/s10822-011-9434-0>.
- (25) Farmer, B. T.; Reitz, A. B. Chapter 11 - Fragment-Based Drug Discovery. In *The Practice of Medicinal Chemistry (Third Edition)*; Wermuth, C. G., Ed.; Academic Press: New York, **2008**; pp 228–243. <https://doi.org/10.1016/B978-0-12-374194-3.00011-1>.
- (26) Schuffenhauer, A.; Ruedisser, S.; Marzinzik, A.; Jahnke, W.; Selzer, P.; Jacoby, E. Library Design for Fragment Based Screening. *Current Topics in Medicinal Chemistry* **5** (8), 751–762.
- (27) Reiher, C. A.; Schuman, D. P.; Simmons, N.; Wolkenberg, S. E. Trends in Hit-to-Lead Optimization Following DNA-Encoded Library Screens. *ACS Med Chem Lett* **2021**, *12* (3), 343–350. <https://doi.org/10.1021/acsmchemlett.0c00615>.
- (28) Gironda-Martínez, A.; Donckele, E. J.; Samain, F.; Neri, D. DNA-Encoded Chemical Libraries: A Comprehensive Review with Successful Stories and Future Challenges. *ACS Pharmacol Transl Sci* **2021**, *4* (4), 1265–1279. <https://doi.org/10.1021/acscptsci.1c00118>.
- (29) Song, M.; Hwang, G. T. DNA-Encoded Library Screening as Core Platform Technology in Drug Discovery: Its Synthetic Method Development and Applications in DEL Synthesis. *J. Med. Chem.* **2020**, *63* (13), 6578–6599. <https://doi.org/10.1021/acs.jmedchem.9b01782>.
- (30) Lavecchia, A.; Giovanni, C. D. Virtual Screening Strategies in Drug Discovery: A Critical Review. *Current Medicinal Chemistry* **20** (23), 2839–2860.
- (31) Brown, D. G. An Analysis of Successful Hit-to-Clinical Candidate Pairs. *J. Med. Chem.* **2023**. <https://doi.org/10.1021/acs.jmedchem.3c00521>.
- (32) Hop, C. E. C. A. Role of ADME Studies in Selecting Drug Candidates: Dependence of ADME Parameters on Physicochemical Properties. In *Encyclopedia of Drug Metabolism and Interactions*; John Wiley & Sons, Ltd, **2012**; pp 1–43. <https://doi.org/10.1002/9780470921920.edm049>.
- (33) Selick, H. E.; Beresford, A. P.; Tarbit, M. H. The Emerging Importance of Predictive ADME Simulation in Drug Discovery. *Drug Discovery Today* **2002**, *7* (2), 109–116. [https://doi.org/10.1016/S1359-6446\(01\)02100-6](https://doi.org/10.1016/S1359-6446(01)02100-6).
- (34) Li, A. P. Screening for Human ADME/Tox Drug Properties in Drug Discovery. *Drug Discovery Today* **2001**, *6* (7), 357–366. [https://doi.org/10.1016/S1359-6446\(01\)01712-3](https://doi.org/10.1016/S1359-6446(01)01712-3).
- (35) Swinney, D. C. Phenotypic vs. Target-Based Drug Discovery for First-in-Class Medicines. *Clinical Pharmacology & Therapeutics* **2013**, *93* (4), 299–301. <https://doi.org/10.1038/clpt.2012.236>.
- (36) *Leading supplier of HTS compounds, building blocks* | *Life Chemicals*. <https://lifechemicals.com/> (accessed 2023-06-06).
- (37) *Home - Enamine*. <https://enamine.net/> (accessed 2023-06-06).
- (38) *ChemBioFrance - Chimiothèque Nationale*. <https://chembiofrance.cn.cnrs.fr/fr/composante/chimiotheque> (accessed 2023-05-21).
- (39) Rudnicki, S. P.; Follen, J. V.; Tolliday, N. J.; Shamu, C. E. Essentials for High-Throughput Screening Operations. In *Chemical Genomics*; Fu, H., Ed.; Cambridge University Press: Cambridge, **2012**; pp 101–107. <https://doi.org/10.1017/CBO9781139021500.012>.
- (40) *Screening Libraries* | *Life Chemicals*. <https://lifechemicals.com/screening-libraries/> (accessed 2023-06-06).
- (41) Goodnow Jr., R. A. The Changing Feasibility and Economics of Chemical Diversity Exploration with DNA-Encoded Combinatorial Approaches. In *A Handbook for DNA-Encoded Chemistry*; John Wiley & Sons, Ltd, **2014**; pp 417–426. <https://doi.org/10.1002/9781118832738.ch18>.
- (42) Burbaum, J. J. Miniaturization Technologies in HTS: How Fast, How Small, How Soon? *Drug Discovery Today* **1998**, *3* (7), 313–322. [https://doi.org/10.1016/S1359-6446\(98\)01203-3](https://doi.org/10.1016/S1359-6446(98)01203-3).

-
- (43) Hibert, M.; Haiech, J. Des gènes aux médicaments : nouveaux défis, nouvelles stratégies. *Med Sci (Paris)* **2000**, *16* (12), 1332. <https://doi.org/10.4267/10608/1586>.
- (44) *ChemBioFrance - Infrastructure de recherche*. <https://chembiofrance.cn.cnr.fr/en/> (accessed 2023-06-11).
- (45) *The NExT Screening Libraries (Pre-plated Copies Available) | Discovery | NExT Resources | NExT*. https://next.cancer.gov/discoveryresources/resources_ndl.htm (accessed 2023-06-09).
- (46) Brennecke, P.; Rasina, D.; Aubi, O.; Herzog, K.; Landskron, J.; Cautain, B.; Vicente, F.; Quintana, J.; Mestres, J.; Stechmann, B.; Ellinger, B.; Brea, J.; Kolanowski, J. L.; Pilarski, R.; Orzaez, M.; Pineda-Lucena, A.; Laraia, L.; Nami, F.; Zielenkiewicz, P.; Paruch, K.; Hansen, E.; von Kries, J. P.; Neuenschwander, M.; Specker, E.; Bartunek, P.; Simova, S.; Leśnikowski, Z.; Krauss, S.; Lehtiö, L.; Bilitewski, U.; Brönstrup, M.; Taskén, K.; Jirgensons, A.; Lickert, H.; Clausen, M. H.; Andersen, J. H.; Vicent, M. J.; Genilloud, O.; Martinez, A.; Nazaré, M.; Fecke, W.; Gribbon, P. EU-OPENSREEN: A Novel Collaborative Approach to Facilitate Chemical Biology. *SLAS Discovery* **2019**, *24* (3), 398–413. <https://doi.org/10.1177/2472555218816276>.
- (47) Parker, C. N.; Schreyer, S. K. Application of Chemoinformatics to High-Throughput Screening. In *Chemoinformatics: Concepts, Methods, and Tools for Drug Discovery*; Bajorath, J., Ed.; Methods in Molecular Biology™; Humana Press: Totowa, NJ, **2004**; pp 85–110. <https://doi.org/10.1385/1-59259-802-1:085>.
- (48) Baybekov, S.; Marcou, G.; Ramos, P.; Saurel, O.; Galzi, J.-L.; Varnek, A. DMSO Solubility Assessment for Fragment-Based Screening. *Molecules* **2021**, *26* (13), 3950. <https://doi.org/10.3390/molecules26133950>.
- (49) Ruggiu, F.; Gizzi, P.; Galzi, J.-L.; Hibert, M.; Haiech, J.; Baskin, I.; Horvath, D.; Marcou, G.; Varnek, A. Quantitative Structure–Property Relationship Modeling: A Valuable Support in High-Throughput Screening Quality Control. *Anal. Chem.* **2014**, *86* (5), 2510–2520. <https://doi.org/10.1021/ac403544k>.
- (50) Rusinko, A.; Farmen, M. W.; Lambert, C. G.; Brown, P. L.; Young, S. S. Analysis of a Large Structure/Biological Activity Data Set Using Recursive Partitioning. *J. Chem. Inf. Comput. Sci.* **1999**, *39* (6), 1017–1026. <https://doi.org/10.1021/ci9903049>.
- (51) van Rhee, A. M.; Stocker, J.; Printzenhoff, D.; Creech, C.; Wagoner, P. K.; Spear, K. L. Retrospective Analysis of an Experimental High-Throughput Screening Data Set by Recursive Partitioning. *J. Comb. Chem.* **2001**, *3* (3), 267–277. <https://doi.org/10.1021/cc0000747>.
- (52) Chang, C.-C.; Lin, C.-J. LIBSVM: A Library for Support Vector Machines. *ACM Trans. Intell. Syst. Technol.* **2011**, *2* (3), 1–27. <https://doi.org/10.1145/1961189.1961199>.
- (53) Fourches, D.; Muratov, E.; Tropsha, A. Trust, But Verify: On the Importance of Chemical Structure Curation in Cheminformatics and QSAR Modeling Research. *J. Chem. Inf. Model.* **2010**, *50* (7), 1189–1204. <https://doi.org/10.1021/ci100176x>.
- (54) *Chemaxon*. <https://chemaxon.com> (accessed 2023-05-21).
- (55) Ruggiu, F.; Marcou, G.; Varnek, A.; Horvath, D. ISIDA Property-Labelled Fragment Descriptors. *Mol. Inf.* **2010**, *29* (12), 855–868. <https://doi.org/10.1002/minf.201000099>.
- (56) Cortes, C.; Vapnik, V. Support-Vector Networks. *Mach Learn* **1995**, *20* (3), 273–297. <https://doi.org/10.1007/BF00994018>.
- (57) Bishop, C. M.; Svensén, M.; Williams, C. K. I. GTM: The Generative Topographic Mapping. *Neural Computation* **1998**, *10* (1), 215–234. <https://doi.org/10.1162/089976698300017953>.
- (58) Kireeva, N.; Baskin, I. 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. *Molecular Informatics* **2012**, *31* (3–4), 301–312. <https://doi.org/10.1002/minf.201100163>.
- (59) Horvath, D.; Marcou, G.; Varnek, A. Generative Topographic Mapping in Drug Design. *Drug Discovery Today: Technologies* **2019**, *32–33*, 99–107. <https://doi.org/10.1016/j.ddtec.2020.06.003>.
- (60) 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* (12), 1087–1108. <https://doi.org/10.1007/s10822-015-9882-z>.
- (61) Russell, S. J.; Norvig, P. *Artificial Intelligence: A Modern Approach*, Fourth edition.; Pearson series in artificial intelligence; Pearson: Hoboken, **2021**.
- (62) *Numerical Recipes: The Art of Scientific Computing*, 3rd ed.; Press, W. H., Ed.; Cambridge University Press: Cambridge, UK; New York, **2007**.
- (63) Mitchell, M. *An Introduction to Genetic Algorithms*, 7. print.; Complex adaptive systems; Cambridge, Mass., **2001**.
- (64) Horvath, D.; Brown, J.; Marcou, G.; Varnek, A. An Evolutionary Optimizer of Libsvm Models. *Challenges* **2014**, *5* (2), 450–472. <https://doi.org/10.3390/challe5020450>.
- (65) Dragos, H.; Gilles, M.; Alexandre, V. Predicting the Predictability: A Unified Approach to the Applicability Domain Problem of QSAR Models. *J. Chem. Inf. Model.* **2009**, *49* (7), 1762–1776. <https://doi.org/10.1021/ci9000579>.
-

- (66) Horvath, D.; Marcou, G.; Varnek, A. A Unified Approach to the Applicability Domain Problem of QSAR Models. *J Cheminform* **2010**, *2* (S1), O6. <https://doi.org/10.1186/1758-2946-2-S1-O6>.
- (67) Varnek, A.; Fourches, D.; Horvath, D.; Klimchuk, O.; Gaudin, C.; Vayer, P.; Solov'ev, V.; Hoonakker, F.; Tetko, I.; Marcou, G. ISIDA - Platform for Virtual Screening Based on Fragment and Pharmacophoric Descriptors. *CAD* **2008**, *4* (3), 191–198. <https://doi.org/10.2174/157340908785747465>.
- (68) Barnett, V.; Lewis, T. *Outliers in Statistical Data*, 3rd ed.; Wiley series in probability and mathematical statistics; Wiley: Chichester; New York, **1994**.
- (69) *Laboratory of Chemoinformatics - Predictor*. <https://chematlas.chimie.unistra.fr/cgi-bin/predictor2.cgi> (accessed 2023-05-24).
- (70) Kevin Oldenburg; Douglas Pooler; Kurt Scudder; Christopher Lipinski; Michele Kelly. High Throughput Sonication: Evaluation for Compound Solubilization. *CCHTS* **2005**, *8* (6), 499–512. <https://doi.org/10.2174/1386207054867364>.
- (71) Di, L.; Kerns, E. H. Chapter 40 - Effects of Properties on Biological Assays. In *Drug-Like Properties (Second Edition)*; Di, L., Kerns, E. H., Eds.; Academic Press: Boston, **2016**; pp 487–496. <https://doi.org/10.1016/B978-0-12-801076-1.00040-X>.
- (72) Cheng, X.; Hochlowski, J.; Tang, H.; Hepp, D.; Beckner, C.; Kantor, S.; Schmitt, R. Studies on Repository Compound Stability in DMSO under Various Conditions. *J Biomol Screen* **2003**, *8* (3), 292–304. <https://doi.org/10.1177/1087057103008003007>.
- (73) Kozikowski, B. A.; Burt, T. M.; Tirey, D. A.; Williams, L. E.; Kuzmak, B. R.; Stanton, D. T.; Morand, K. L.; Nelson, S. L. The Effect of Freeze/Thaw Cycles on the Stability of Compounds in DMSO. *SLAS Discovery* **2003**, *8* (2), 210–215. <https://doi.org/10.1177/1087057103252618>.
- (74) Lipinski, C. Solubility in the Design of Combinatorial Libraries. In *Analysis and Purification Methods in Combinatorial Chemistry*; John Wiley & Sons, Ltd, **2003**; pp 407–434. <https://doi.org/10.1002/0471531979.ch16>.
- (75) Hoefer, M.; Zbinden, P. The Evolution of Microarrayed Compound Screening. *Drug Discovery Today* **2004**, *9* (8), 358–365. [https://doi.org/10.1016/S1359-6446\(04\)03037-5](https://doi.org/10.1016/S1359-6446(04)03037-5).
- (76) Baybekov, Shamkhal; Marcou, Gilles; Ramos, Pascal; Saurel, Olivier; Galzi, Jean-Luc; Varnek, Alexandre. DMSO Solubility Assessment for Fragment-Based Screening, **2021**. <https://doi.org/10.5281/ZENODO.4767511>.
- (77) Di, L.; Fish, P. V.; Mano, T. Bridging Solubility between Drug Discovery and Development. *Drug Discovery Today* **2012**, *17* (9–10), 486–495. <https://doi.org/10.1016/j.drudis.2011.11.007>.
- (78) Stuart, M.; Box, K. Chasing Equilibrium: Measuring the Intrinsic Solubility of Weak Acids and Bases. *Anal. Chem.* **2005**, *77* (4), 983–990. <https://doi.org/10.1021/ac048767n>.
- (79) Hasselbalch, K. A. Calculation of Blood PH Based on the Free and Bound Carbonic Acid, and Oxygen Binding of Blood as Function of PH. *Die Biochem. Z* **1916**, *78*, 112–144.
- (80) Bergström, C. A. S.; Luthman, K.; Artursson, P. Accuracy of Calculated PH-Dependent Aqueous Drug Solubility. *European Journal of Pharmaceutical Sciences* **2004**, *22* (5), 387–398. <https://doi.org/10.1016/j.ejps.2004.04.006>.
- (81) Po, H. N.; Senozan, N. M. The Henderson-Hasselbalch Equation: Its History and Limitations. *J. Chem. Educ.* **2001**, *78* (11), 1499. <https://doi.org/10.1021/ed078p1499>.
- (82) Di, L.; Kerns, E. H. Chapter 7 - Solubility. In *Drug-Like Properties (Second Edition)*; Di, L., Kerns, E. H., Eds.; Academic Press: Boston, **2016**; pp 61–93. <https://doi.org/10.1016/B978-0-12-801076-1.00007-1>.
- (83) Baybekov, S.; Llompart, P.; Marcou, G.; Gizzi, P.; Galzi, J.-L.; Ramos, P.; Saurel, O.; Bourban, C.; Minoletti, C.; Varnek, A. Données de Réplication Pour : Kinetic Solubility: Experimental and Machine-Learning Modeling Perspectives, **2023**. <https://doi.org/10.57745/ZWS0WC>.
- (84) Llompart, P.; Minoletti, C.; Baybekov, S.; Horvath, D.; Marcou, G.; Varnek, A. Données de Réplication Pour : Towards the Improvement of Thermodynamic Solubility Prediction – a Review, **2023**. <https://doi.org/10.57745/CZVZIA>.
- (85) European Parliament and Council. *Regulation (EC) No 1223/2009 of the European Parliament and of the Council of 30 November 2009 on Cosmetic Products (Recast) (Text with EEA Relevance)*; **2009**; Vol. 342. <http://data.europa.eu/eli/reg/2009/1223/oj/eng> (accessed 2022-10-27).
- (86) Brown, T. N.; Armitage, J. M.; Egeghy, P.; Kircanski, I.; Arnot, J. A. Dermal Permeation Data and Models for the Prioritization and Screening-Level Exposure Assessment of Organic Chemicals. *Environment International* **2016**, *94*, 424–435. <https://doi.org/10.1016/j.envint.2016.05.025>.
- (87) Potts, R. O.; Guy, R. H. Predicting Skin Permeability. *Pharm Res* **1992**, *9* (5), 663–669. <https://doi.org/10.1023/A:1015810312465>.

-
- (88) Moss, G. P.; Cronin, M. T. D. Quantitative Structure–Permeability Relationships for Percutaneous Absorption: Re-Analysis of Steroid Data. *International Journal of Pharmaceutics* **2002**, *238* (1), 105–109. [https://doi.org/10.1016/S0378-5173\(02\)00057-1](https://doi.org/10.1016/S0378-5173(02)00057-1).
- (89) Waters, L. J.; Quah, X. L. Predicting Skin Permeability Using HuskinDB. *Sci Data* **2022**, *9* (1), 584. <https://doi.org/10.1038/s41597-022-01698-4>.
- (90) Schreiber, S. L.; Kotz, J. D.; Li, M.; Aubé, J.; Austin, C. P.; Reed, J. C.; Rosen, H.; White, E. L.; Sklar, L. A.; Lindsley, C. W.; Alexander, B. R.; Bittker, J. A.; Clemons, P. A.; de Souza, A.; Foley, M. A.; Palmer, M.; Shamji, A. F.; Wawer, M. J.; McManus, O.; Wu, M.; Zou, B.; Yu, H.; Golden, J. E.; Schoenen, F. J.; Simeonov, A.; Jadhav, A.; Jackson, M. R.; Pinkerton, A. B.; Chung, T. D. Y.; Griffin, P. R.; Cravatt, B. F.; Hodder, P. S.; Roush, W. R.; Roberts, E.; Chung, D.-H.; Jonsson, C. B.; Noah, J. W.; Severson, R. E.; Ananthan, S.; Edwards, B.; Oprea, T. I.; Conn, P. J.; Hopkins, C. R.; Wood, M. R.; Stauffer, S. R.; Emmitte, K. A.; Brady, L. S.; Driscoll, J.; Li, I. Y.; Loomis, C. R.; Margolis, R. N.; Michelotti, E.; Perry, M. E.; Pillai, A.; Yao, Y. Advancing Biological Understanding and Therapeutics Discovery with Small-Molecule Probes. *Cell* **2015**, *161* (6), 1252–1265. <https://doi.org/10.1016/j.cell.2015.05.023>.
- (91) Workman, P.; Collins, I. Probing the Probes: Fitness Factors For Small Molecule Tools. *Chemistry & Biology* **2010**, *17* (6), 561–577. <https://doi.org/10.1016/j.chembiol.2010.05.013>.
- (92) Arrowsmith, C. H.; Audia, J. E.; Austin, C.; Baell, J.; Bennett, J.; Blagg, J.; Bountra, C.; Brennan, P. E.; Brown, P. J.; Bunnage, M. E.; Buser-Doepner, C.; Campbell, R. M.; Carter, A. J.; Cohen, P.; Copeland, R. A.; Cravatt, B.; Dahlin, J. L.; Dhanak, D.; Edwards, A. M.; Frederiksen, M.; Frye, S. V.; Gray, N.; Grimshaw, C. E.; Hepworth, D.; Howe, T.; Huber, K. V. M.; Jin, J.; Knapp, S.; Kotz, J. D.; Kruger, R. G.; Lowe, D.; Mader, M. M.; Marsden, B.; Mueller-Fahrnow, A.; Müller, S.; O’Hagan, R. C.; Overington, J. P.; Owen, D. R.; Rosenberg, S. H.; Ross, R.; Roth, B.; Schapira, M.; Schreiber, S. L.; Shoichet, B.; Sundström, M.; Superti-Furga, G.; Taunton, J.; Toledo-Sherman, L.; Walpole, C.; Walters, M. A.; Willson, T. M.; Workman, P.; Young, R. N.; Zuercher, W. J. The Promise and Peril of Chemical Probes. *Nat Chem Biol* **2015**, *11* (8), 536–541. <https://doi.org/10.1038/nchembio.1867>.
- (93) Rayevsky, A. V.; Poturai, A. S.; Kravets, I. O.; Pashenko, A. E.; Borisova, T. A.; Tolstanova, G. M.; Volochnyuk, D. M.; Borysko, P. O.; Vadzyuk, O. B.; Aliksieieva, D. O.; Zabolotna, Y.; Klimchuk, O.; Horvath, D.; Marcou, G.; Ryabukhin, S. V.; Varnek, A. In Vitro Evaluation of In Silico Screening Approaches in Search for Selective ACE2 Binding Chemical Probes. *Molecules* **2022**, *27* (17), 5400. <https://doi.org/10.3390/molecules27175400>.
- (94) Zabolotna, Y.; Volochnyuk, D. M.; Ryabukhin, S. V.; Gavrylenko, K.; Horvath, D.; Klimchuk, O.; Oksiuta, O.; Marcou, G.; Varnek, A. SynthI: A New Open-Source Tool for Synthron-Based Library Design. *J. Chem. Inf. Model.* **2022**, *62* (9), 2151–2163. <https://doi.org/10.1021/acs.jcim.1c00754>.
- (95) Basu, A.; Sarkar, A.; Maulik, U. Molecular Docking Study of Potential Phytochemicals and Their Effects on the Complex of SARS-CoV2 Spike Protein and Human ACE2. *Sci Rep* **2020**, *10* (1), 17699. <https://doi.org/10.1038/s41598-020-74715-4>.
- (96) Rola, M.; Krassowski, J.; Górska, J.; Grobelna, A.; Plonka, W.; Paneth, A.; Paneth, P. Machine Learning Augmented Docking Studies of Aminothiouras at the SARS-CoV-2—ACE2 Interface. *PLOS ONE* **2021**, *16* (9), e0256834. <https://doi.org/10.1371/journal.pone.0256834>.
- (97) Zarezade, V.; Rezaei, H.; Shakerinezhad, G.; Safavi, A.; Nazeri, Z.; Veisi, A.; Azadbakht, O.; Hatami, M.; Sabaghan, M.; Shajirat, Z. The Identification of Novel Inhibitors of Human Angiotensin-Converting Enzyme 2 and Main Protease of Sars-Cov-2: A Combination of in Silico Methods for Treatment of COVID-19. *Journal of Molecular Structure* **2021**, *1237*, 130409. <https://doi.org/10.1016/j.molstruc.2021.130409>.
- (98) Hochuli, J. E.; Jain, S.; Melo-Filho, C.; Sessions, Z. L.; Bobrowski, T.; Choe, J.; Zheng, J.; Eastman, R.; Talley, D. C.; Rai, G.; Simeonov, A.; Tropsha, A.; Muratov, E. N.; Baljinnyam, B.; Zakharov, A. V. Allosteric Binders of ACE2 Are Promising Anti-SARS-CoV-2 Agents. *ACS Pharmacol. Transl. Sci.* **2022**, *5* (7), 468–478. <https://doi.org/10.1021/acspstsci.2c00049>.
- (99) Cohen, W. W. Fast Effective Rule Induction. In *Machine Learning Proceedings 1995*; Prieditis, A., Russell, S., Eds.; Morgan Kaufmann: San Francisco (CA), **1995**; pp 115–123. <https://doi.org/10.1016/B978-1-55860-377-6.50023-2>.
- (100) *Data Mining: Practical Machine Learning Tools and Techniques*, Fourth Edition.; Witten, I. H., Witten, I. H., Eds.; Elsevier: Amsterdam, **2017**.
- (101) RCSB Protein Data Bank. RCSB PDB - 1R4L: Inhibitor Bound Human Angiotensin Converting Enzyme-Related Carboxypeptidase (ACE2). <https://www.rcsb.org/structure/1R4L> (accessed 2023-08-26).
- (102) Wolber, G.; Langer, T. LigandScout: 3-D Pharmacophores Derived from Protein-Bound Ligands and Their Use as Virtual Screening Filters. *J. Chem. Inf. Model.* **2005**, *45* (1), 160–169. <https://doi.org/10.1021/ci049885e>.
-

- (103) Korb, O.; Stützle, T.; Exner, T. E. An Ant Colony Optimization Approach to Flexible Protein–Ligand Docking. *Swarm Intell* **2007**, *1* (2), 115–134. <https://doi.org/10.1007/s11721-007-0006-9>.
- (104) Korb, O.; Stützle, T.; Exner, T. E. Empirical Scoring Functions for Advanced Protein–Ligand Docking with PLANTS. *J. Chem. Inf. Model.* **2009**, *49* (1), 84–96. <https://doi.org/10.1021/ci800298z>.
- (105) Crouch, S. P. M.; Slater, K. J. High-Throughput Cytotoxicity Screening: Hit and Miss. *Drug Discovery Today* **2001**, *6*, 48–53. [https://doi.org/10.1016/S1359-6446\(01\)00151-9](https://doi.org/10.1016/S1359-6446(01)00151-9).
- (106) Kerns, E. H.; Di, L. Chapter 26 - Permeability Methods. In *Drug-like Properties: Concepts, Structure Design and Methods*; Kerns, E. H., Di, L., Eds.; Academic Press: San Diego, **2008**; pp 287–298. <https://doi.org/10.1016/B978-012369520-8.50027-9>.
- (107) Reker, D.; Schneider, P.; Schneider, G.; Brown, J. Active Learning for Computational Chemogenomics. *Future Medicinal Chemistry* **2017**, *9* (4), 381–402. <https://doi.org/10.4155/fmc-2016-0197>.
- (108) P. Mazanetz, M.; J. Marmon, R.; B. T. Reisser, C.; Morao, I. Drug Discovery Applications for KNIME: An Open Source Data Mining Platform. *Current Topics in Medicinal Chemistry* **2012**, *12* (18), 1965–1979. <https://doi.org/10.2174/156802612804910331>.
- (109) Stöter, M.; Niederlein, A.; Barsacchi, R.; Meyenhofer, F.; Brandl, H.; Bickle, M. CellProfiler and KNIME: Open Source Tools for High Content Screening. In *Target Identification and Validation in Drug Discovery: Methods and Protocols*; Moll, J., Colombo, R., Eds.; Methods in Molecular Biology; Humana Press: Totowa, NJ, **2013**; pp 105–122. https://doi.org/10.1007/978-1-62703-311-4_8.
- (110) Strobelt, H.; Bertini, E.; Braun, J.; Deussen, O.; Groth, U.; Mayer, T. U.; Merhof, D. HiTSEE KNIME: A Visualization Tool for Hit Selection and Analysis in High-Throughput Screening Experiments for the KNIME Platform. *BMC Bioinformatics* **2012**, *13* (S8), S4. <https://doi.org/10.1186/1471-2105-13-S8-S4>.
- (111) Macmillan, D. S.; Chilton, M. L.; Gao, Y.; Kern, P. S.; Schneider, S. N. How to Resolve Inconclusive Predictions from Defined Approaches for Skin Sensitisation in OECD Guideline No. 497. *Regulatory Toxicology and Pharmacology* **2022**, *135*, 105248. <https://doi.org/10.1016/j.yrtph.2022.105248>.
- (112) Li, H. L. Improved Defined Approaches for Predicting Skin Sensitization Hazard and Potency in Humans. *ALTEX* **2019**. <https://doi.org/10.14573/altex.1809191>.
- (113) *EU Reference Laboratory for alternatives to animal testing (EURL ECVAM)*. https://joint-research-centre.ec.europa.eu/eu-reference-laboratory-alternatives-animal-testing-eurl-ecvam_en (accessed 2023-06-13).
- (114) Zuang, V.; Daskalopoulos, E.; Berggren, E.; Batista, L. S.; Bopp, S.; Carpi, D.; Casati, S.; Corvi, R.; Cusinato, A.; Deceuninck, P.; Dura, A.; Franco, A.; Gastaldello, A.; Gribaldo, L.; Holloway, M.; Katsanou, E.; Langezaal, I.; Morath, S.; Munn, S.; Prieto, P. M. D. P.; Piergiovanni, M.; Whelan, M.; Wittwehr, C.; Worth, A.; Viegas, B. J. F. *Non-animal Methods in Science and Regulation*. JRC Publications Repository. <https://doi.org/10.2760/500414>.

Appendix

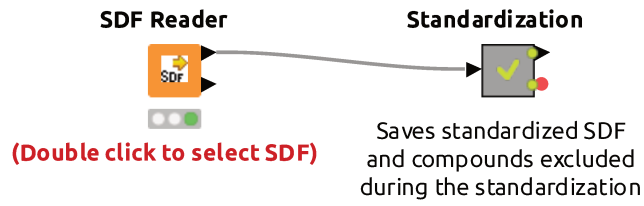


Figure 21. Screenshot of the "1_standardization" KNIME workflow.

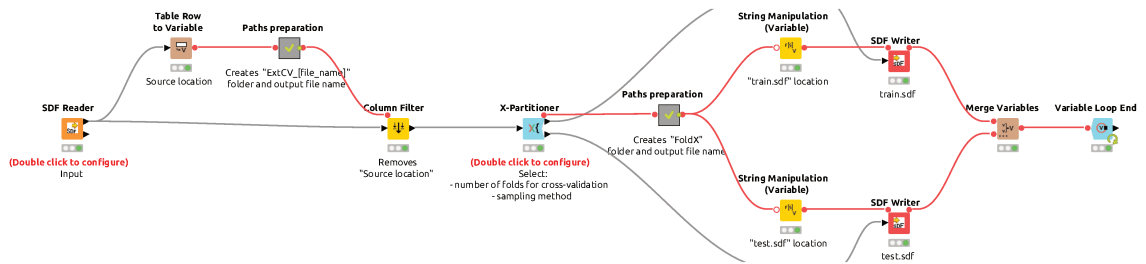


Figure 22. Screenshot of the "2_ExtCV_data_partitioning" KNIME workflow.

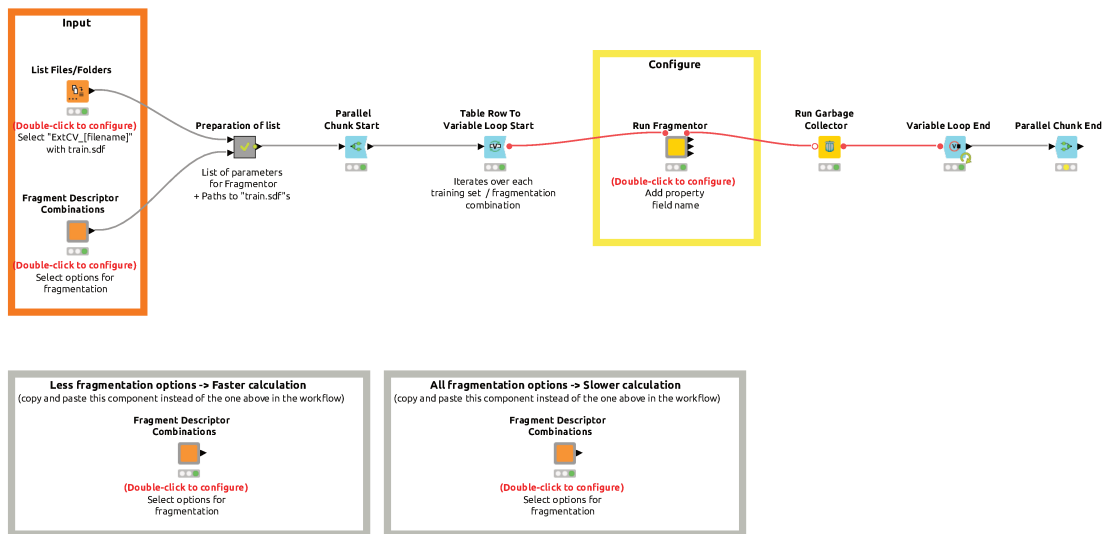


Figure 23. Screenshot of the "3_ExtCV_descriptor_calculation" KNIME workflow.

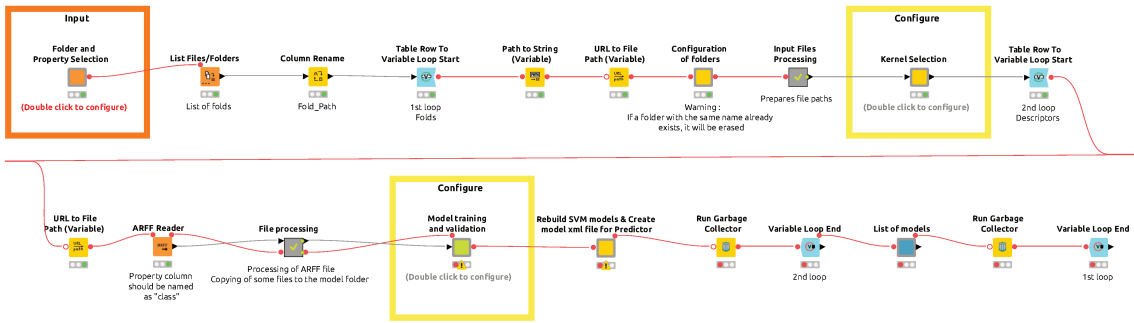


Figure 24. Screenshot of the "4_ExtCV_modeling_CLS" KNIME workflow.

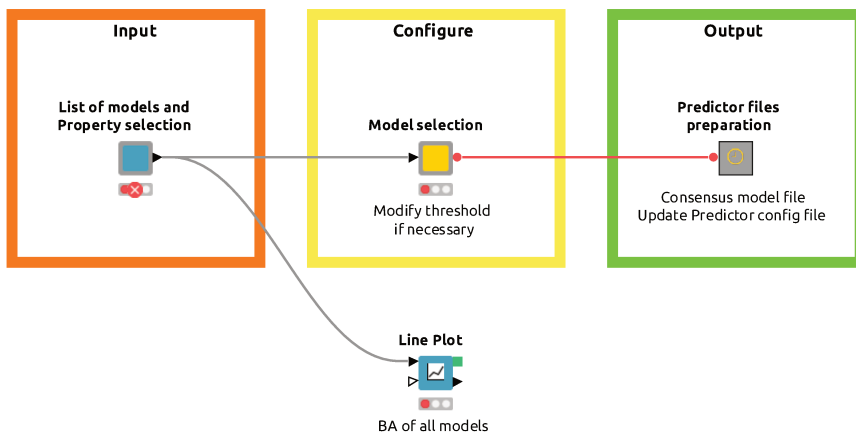


Figure 25. Screenshot of the "5_ExtCV_consensus_preparation_CLS" KNIME workflow.

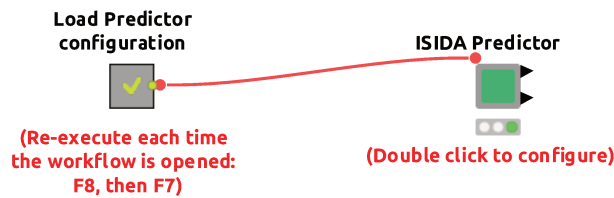


Figure 26. Screenshot of the "6_ExtCV_application" KNIME workflow.

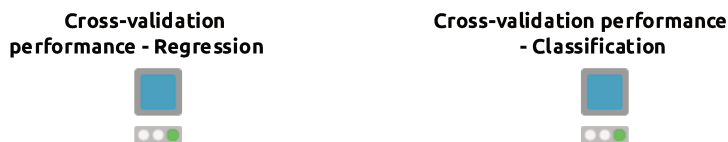


Figure 27. Screenshot of the "7_ExtCV_evaluation" KNIME workflow.

List of Figures

Figure 1. Vue d'ensemble du processus de criblage et de l'application de chémoinformatique à chaque étape du processus.....	2
Figure 2. Division de la plage de solubilité en catégories fixées par les seuils pour le criblage basé sur les fragments (1 mM) et pour la formulation de solutions mères (10 mM). Les étiquettes "Soluble" et "Insoluble" coïncident pour les solubilités supérieures à 10 mM et inférieures à 1 mM. Toutefois, dans la plage de 1 à 10 mM, les composés sont considérés comme solubles selon la définition du FBS, mais insolubles selon la définition d'une solution mère.....	4
Figure 3. Types d'essais de mesure de la solubilité. HTS, MTS et LTS désignent respectivement un criblage à haut, moyen et bas débit.....	5
Figure 4. Voie d'expression des effets indésirables (AOP) du processus de sensibilisation cutanée. Une liste non exhaustive d'essais bien connus décrivant chaque événement clef (KE) est donnée sous leur KE respectif.....	7
Figure 5. Processus de collecte et de filtrage des données de la nouvelle base de données SkinPiX. "SC" signifie stratum corneum. Le processus suit deux étapes principales. Tout d'abord, les publications scientifiques pertinentes ont été recherchées dans PubMed. Ensuite, les données sur la perméabilité de la peau ont été extraites avec leurs métadonnées. Seules les données répondant à des critères spécifiques ont été conservées, comme illustré.....	8
Figure 6. Aperçu des étapes du criblage virtuel pour identifier les inhibiteurs sélectifs de l'ACE2.	9
Figure 7. Distribution of lead generation strategies used in 156 successful hit-to-clinical campaigns. The figure is adapted from Brown ³¹	17
Figure 8. Simplified overview of a screening campaign: (1) chemical library design, (2) stock preparation, (3) sample preparation, (4) performing the test, (5) data acquisition, (6) data analysis.....	18
Figure 9. Overview of screening workflow and application of chemoinformatics methods at each step of the workflow.	22
Figure 10. Example of fragmentation to ISIDA substructural molecular fragments. Stars annotated different carbon atoms. Circles highlight atom centers. The number of occurrences is given below each fragment. In triplets, the number between each atom pair indicates topological distance, or number of bonds between two atoms.....	26
Figure 11. General modeling workflow. ML stands for machine learning.....	32
Figure 12. Types of aqueous solubility and applied solubility measurement assay. HTS, MTS and LTS stand for high-, medium- and low-throughput screening, respectively.....	47
Figure 13. Data processing of the HuskinDB.....	103
Figure 14. 5-fold cross-validation performance of "HuskinDB model" and "merged model".....	104
Figure 15. Performance of "HuskinDB model" on SkinPiX test set.....	104
Figure 16. GTM class landscape showing the distribution of the HuskinDB training dataset (blue) and the SkinPiX test set (red). Yellow and green areas are populated with compounds of both datasets. The structural motifs found in certain zones are displayed.	105
Figure 17. Overview of virtual screening steps to identify selective ACE2 inhibitors.	108
Figure 18. Classification rules derived by the JRip rule-based algorithm for ACE2 inhibition.	112
Figure 19. Structure-based pharmacophore model. Red spheres – H-bond acceptor; green spheres – H-bond donor; yellow spheres – hydrophobic regions.....	112
Figure 20. Screenshots showing example of request for ISIDA Predictor web service. Image A shows the ISIDA Predictor configuration page, where a user can select "Activity"/"PhysProp" general kind of property and then choose the model of interest. Image B illustrates an output of ISIDA Predictor. Color code of prediction confidence is as follows: green – optimal; blue – good; orange – average; red – unreliable.	117
Figure 21. Screenshot of the "1_standardization" KNIME workflow.	125
Figure 22. Screenshot of the "2_ExtCV_data_partitioning" KNIME workflow.	125
Figure 23. Screenshot of the "3_ExtCV_descriptor_calculation" KNIME workflow.	125
Figure 24. Screenshot of the "4_ExtCV_modeling_CLS" KNIME workflow.	126
Figure 25. Screenshot of the "5_ExtCV_consensus_preparation_CLS" KNIME workflow.....	126
Figure 26. Screenshot of the "6_ExtCV_application" KNIME workflow.	126
Figure 27. Screenshot of the "7_ExtCV_evaluation" KNIME workflow.	126
Figure 28. Screenshot of the "8_final_consensus_preparation_CLS" KNIME workflow.	127

List of Tables

Tableau 1. La liste des procédures automatisées KNIME développées.....	10
Tableau 2. La liste des modèles QSAR/QSPR développés, la taille de leurs ensembles d'apprentissage et les valeurs de performance. VC 5-fois - la validation croisée 5-fois ; BA - la précision balancée ; RMSE - racine de l'erreur quadratique moyenne. Remarque concernant la BA : BA = 0,5 - prédiction aléatoire ; BA = 1 - prédiction parfaite. Remarque concernant le RMSE : plus le RMSE est petit, meilleur est le modèle. Entre parenthèses, la taille de l'ensemble de test est indiquée.....	11
Tableau 3. La liste des modèles développés et les moyens d'y accéder. Tous les modèles (sauf la solubilité aqueuse thermodynamique) sont accessibles sur la page web https://chematlas.chimie.unistra.fr/cgi-bin/predictor2.cgi	12
Table 4. Screening campaign types categorized by measurement rate, approach, and assay type.....	16
Table 5. Cost estimates per sample (US\$) for preparation of compounds for HTS campaigns. The table is adapted from Goodnow ⁴¹	19
Table 6. A list of developed KNIME workflows.....	33
Table 7. Training set sizes of ACE2, ACE and NEP models.....	109
Table 8. Performance of ACE2, ACE and NEP consensus QSAR models, and their constituting models. Fragmentation types: I – sequence; II – atom-centered; III – triplet; A – atom; B – bond; R – fragment of fixed length; P – “Atom Pairs” option; AP – “Do All Ways” option. BA _{5-CV} – balanced accuracy on 5-fold cross-validation.....	110
Table 9. The list of published data and links to access them.....	116
Table 10. The list of developed QSAR/QSPR models, their training set sizes and predictive performance values. The size of the test set is indicated in brackets. BA – balanced accuracy; CV – cross-validation; RMSE – root mean-squared error.....	116
Table 11. The list of developed models and how to access them. All models (except thermodynamic aqueous solubility) are available on the https://chematlas.chimie.unistra.fr/cgi-bin/predictor2.cgi web page. The models can be accessed by first selecting "General kind of property" and then "Property to model"....	116

Modélisation QSAR et QSPR de propriétés d'intérêt pour le criblage et la sécurité des composés

Résumé

Cette thèse concerne le développement et la mise en œuvre d'outils de chémoinformatique en support de campagnes de criblage de composés. Les sujets couverts sont : les étapes de sélection de composés, d'évaluation de l'intégrité des solutions de stock, de contrôle de la qualité des données expérimentales, de développement de modèles prédictifs et d'annotation des bibliothèques de criblage. Les propriétés d'intérêt incluent la solubilité de composés de type fragment dans le DMSO, la solubilité aqueuse, la sensibilisation cutanée, la perméabilité cutanée et la liaison à l'enzyme de conversion de l'angiotensine (ACE2) pour la conception de sondes biologiques. Les modèles de relation structure-activité/propriété quantitative (QSAR/QSPR) sont disponibles publiquement et des outils conviviaux développés dans la KNIME Analytics Platform fournissent un soutien précieux aux chercheurs sans nécessité de compétences en programmation. L'intégration des outils de chémoinformatique développés offre une approche efficace pour améliorer les résultats du criblage et maximiser l'efficacité.

Mots-clés : QSAR, QSPR, criblage, solubilité, perméabilité cutanée, sensibilisation cutanée, inhibition de l'ACE2

Résumé en anglais

This thesis concerns the development and implementation of chemoinformatics tools to support compound screening campaigns. It covers the following topics: steps for compound selection, assessment of stock solution integrity, quality control of experimental data, development of predictive models, and annotation of screening libraries. The properties of interest include solubility of fragment-like compounds in DMSO, aqueous solubility, skin sensitization, skin permeability, and binding to angiotensin-converting enzyme (ACE2) for the design of biological probes. The publicly available quantitative structure-activity/property relationship (QSAR/QSPR) models and user-friendly tools developed in KNIME Analytics Platform provide valuable support to researchers without the need for coding expertise. The integration of the developed chemoinformatics tools offers an efficient approach to improving screening outcomes and maximizing efficiency.

Keywords : QSAR, QSPR, screening, solubility, skin permeability, skin sensitization, ACE2 inhibition