UNIVERSITÉ DE STRASBOURG

ÉCOLE DOCTORALE DES SCIENCES CHIMIQUES

UMR 7140

THÈSE présentée par :

Shilva KAYASTHA

soutenue le : **19 septembre 2017**

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

Discipline/ Spécialité : **Chimie/Chémoinformatique**

New methods of multiscale chemical space analysis: visualization of structure-activity relationships and structural pattern extraction

THÈSE dirigée par :

M. VARNEK Alexandre	Professeur, Université de Strasbourg
M. BAJORATH Jürgen	Professeur, Université de Bonn, Allemagne

RAPPORTEURS :

M. REYMOND Jean-Louis	Professeur, Université de Bern, Suisse
M. PETITJEAN Michel	CR1 CNRS, HDR, Université Paris Diderot

AUTRES MEMBRES DU JURY :

M. ROGNAN Didier	DR1 CNRS, Université de Strasbourg
M. HORVATH Dragos	DR2 CNRS, Université de Strasbourg

UNIVERSITÉ DE STRASBOURG

Shilva Kayastha

EDSC
École Doctorale des
Sciences Chimiques

Nouvelles méthodes d'analyse multi-échelle de l'espace chimique: visualisation de relations structure-activité et l'extraction des motifs structuraux

Résumé

Cette thèse est dédiée à l’analyse systématique de l’espace chimique, et des relations structure-activité (SAR) en particulier. L’ouvrage présente des nouveaux protocoles d’analyse combinant des méthodes classiques et originales, dans le but d’analyser les SAR à l’échelle globale ainsi que locale.

L’analyse globale des espaces chimiques repose sur la recherche des motifs structuraux privilégiés par cartographie topographique générative (GTM), ainsi que par analyse classique des « châssis » moléculaires. La cartographie a été ensuite couplée avec l’analyse de réseaux chimiques (CSN), permettant une transition de la vue globale vers l’analyse locale de SAR. L’optimisation mutiobjectif des propriétés de potentiels médicaments a été adressé par la méthode « star coordinates ».

L’analyse locale des SAR inclut des nouvelles stratégies pour prédire les discontinuités dans le paysage structure-activité biologique, et une étude de l’impact de la structure sur l’ionisation des molécules. Des matrices SAR ont servi pour monitorer le progrès dans l’optimisation de nouveaux principes actifs

Mots clefs : visualisation de l’espace chimique, SAR, GTM, CSN

Résumé en anglais

This thesis presents studies devoted to aid in systematic analysis of chemical spaces, focusing on mining and visualization of structure-activity relationships (SARs). It reports some new analysis protocols, combining both existing and on-purpose developed novel methodology to address both large-scale and local SAR analysis.

Large-scale analysis featured both generative topographic mapping (GTM)-based extraction of privileged structural motifs and scaffold analysis. GTM was combined with chemical space network (CSN) to develop a visualization tool providing global-local views of SAR in large data sets. We also introduce star coordinates (STC) to visualize multi-property space and prioritize drug-like subspaces.

Local SAR monitoring includes new strategies to predict activity cliffs using support vector machine models and a study of structural modifications on ionization state of compounds. The SAR matrix methodology was applied to objectively evaluate SAR progression during lead optimization.

Key words: chemical space visualization, SAR, GTM, CSN

To my beloved parents and sister,
for all the love and endless support.

Acknowledgements

To my group at the University of Bonn:

First, I would like to express my gratitude to Prof. Dr. Jürgen Bajorath for his valuable guidance, supervision and profound knowledge that constantly inspired me to work towards my scientific development. Further, his continuous support and thoughtful advices greatly helped me to stay motivated even during the most difficult times along my PhD journey. Thank you for everything.

My heartfelt thanks goes to Dilyana Dimova for her great encouragement, support, and suggestions, both in scientific and personal matters. I will forever be grateful for the wonderful friendship and good times we shared. I would also like to thank Antonio de la Vega de León for being an excellent team partner in all our collaborative projects. Thank you for being always helpful and cheerful that made work more fun. Special thanks to Dagmar Stumpfe and Ye Hu for their help and advice on many occasions as well as their warm and friendly attitude. It is a pleasure to acknowledge all the group members of Prof. Bajorath for creating an outstanding environment at work from where I have a lot of fond memories.

To my group at the University of Strasbourg:

My sincere thanks goes to my supervisor Prof. Dr. Alexandre Varnek for giving me an opportunity to be part of exciting joint projects between the two universities. His optimism and enthusiasm were essential to bring this collaborative effort into fruition. Discussions with him and advices given by him were very helpful.

I am deeply grateful to Dragos Horvath for having patience to answer all my questions and making me familiar with the new tools. Constructive comments and suggestions given by him helped materialize many studies in this thesis. I am grateful to Fanny Bonachera for her help with translation of the thesis abstract to French. I am indebted to Olga Klimchuck for her support throughout all the administrative processes.

I would like to acknowledge Region Alsace for their generous support in funding my PhD studies.

I thank Prof. Dr. Jean-Louis Reymond,Dr. Michel Petitjean, and Dr. Didier Rognan for agreeing to be part of my thesis defense committee.

Lastly, but most importantly, I owe my deepest gratitude to my beloved family members for their immeasurable support and encouragement from early on in my life. The values instilled by you has inspired me to achieve my goals and I cannot thank you enough for that. Special thanks goes to Reagon Karki for his patience and uplifting spirit during stressful moments.

Contents

Résumé de these en français

Introduction

Cette thèse présente plusieurs études consacrées au développement de nouvelles approches computationnelles pour l'analyse systématique de données et à la visualisation d'espaces chimiques. Les projets discutés ici ont pour but de traiter différents challenges rencontrés lors de l'optimisation de composés ainsi que lors du processus de mise au point de médicaments. De plus, des analyses à grande échelle de jeux de données de composés ont été menées pour déceler des motifs structuraux, ce qui présente d'importantes implications pour concevoir de potentiels nouveaux médicaments. Sept études sont discutées dans cette thèse, qui inclut trois domaines principaux : la fouille de données, la visualisation et la prédiction. Plusieurs outils ont été utilisés pour l'analyse de l'espace chimique, comme les matrices de relations structure-activité (SAR matrix - SARM), la visualisation par star coordinates (STC), les GTMs (generative topographic maps – ISIDA-GTM), ainsi que les réseaux d'espace chimique (chemical space networks – CSNs). La méthode STC a été développée et utilisée ici pour la première fois pour visualiser la distribution des données dans l'espace chimique. Bien que les autres outils aient été déjà présentés dans de précédentes études, nous avons démontré leur application dans différents aspects de la découverte et de la conception de médicaments par ordinateur.

Cette thèse est organisée en 9 chapitres. Le chapitre 1 introduit les concepts clés et les méthodologies. Le chapitre 2 explore la diversité structurale et la distribution de l'activité des squelettes structuraux (scaffolds) de l'univers actuel des composés bioactifs. Dans le chapitre 3, nous démontrons l'application des GTMs pour détecter les motifs structuraux privilégiés dans des classes de composés. Un outil de visualisation de relations structure-activités à deux couches, combinant les GTMs et les CSNs, est présenté dans le chapitre 4. L'application du GTM-CSN pour détecter des schémas de relations structure-activité intéressants dans de très grands jeux de données, y compris le jeu de données de criblage anti-malaria comprenant plus de 13000 composés, est présentée. Dans le chapitre 5, la prédiction avec succès de pics d'activité (activity cliffs) via l'utilisation de la classification par vecteurs de supports et modèles de régression est discutée. Une analyse à grande échelle de la distribution des états d'ionisation des composés bioactifs est présentée dans le chapitre 6. Le chapitre 7 introduit un outil pour la visualisation des composés dans un paysage à multi-propriétés. Un outil de diagnostic, utilisé pour surveiller la progression des relations structure-activité (structure-activity relationship – SAR) lors de l'optimisation de composés prometteurs, est introduit dans le chapitre 8. Enfin, le chapitre 9 comprend un résumé ainsi que des conclusions.

Distribution de la diversité structurale et de l'échelle d'activité des scaffolds dans l'univers actuel des composés bioactifs.

Le concept de squelette structural (scaffold) est appliqué en chimie médicinale pour décrire la structure de base des composés actifs et est resté assez populaire au cours des années. Il a été utilisé dans des applications variées, comme par exemple pour chercher les sous-structures privilégiées[1] dans une famille de cibles données, ou pour détecter des occurrences de scaffold-hopping[2] (sortir d'un modèle structural donné) dans des jeux de données de composés. Etant donnée l'augmentation du nombre de nouveaux composés ajoutés dans l'univers actuel des composés bioactifs, l'organisation systématique et l'analyse de leurs motifs structuraux est devenue plus complexe. Une étude exhaustive a été rapportée dans ces travaux, avec comme but de déterminer de façon systématique la diversité structurale des composés, de façon globale mais aussi au niveau des cibles individuelles. Dans notre analyse, nous avons utilisé la définition du motif structural décrite par Bemis-Murcko,[3] et un niveau d'abstraction supplémentaire a été obtenu en générant des squelettes cycliques (cyclic skeletons – CSKs). Les scaffolds et les CSKs ont été extraits de façon systématique à partir des composés dont les données d'activité présentent une confiance élevée dans la base de données ChEMBL version 18. Les scaffolds et les CSKs ont été assignés à des intervalles d'activité basés sur la distribution de l'intervalle d'activité des composés représentés par ces mêmes motifs structuraux et CSKs. Des scaffolds structuralement différents sont fréquemment apparus dans nos analyses. Ils représentaient des composés hautement actifs contre de nombreuses cibles pharmaceutiques. Par conséquent, il est probable que des séries de composés structuralement distincts puissent être développées afin de rechercher des candidats médicaments alternatifs pour ce type de cibles.

Analyse des états d'ionisation de composés bioactifs, basée sur la structure de la cible

L'état de charge des petites molécules est une des propriétés les plus importantes, qui détermine leur activité biologique et leur action en tant que médicament. L'état d'ionisation des composés bioactifs peut changer en fonction du pH dans différents environnements in vivo, ce qui, par conséquent, affecte leurs profils pharmacologiques. Cette étude se concentre sur les relations entre les états d'ionisation des composés, la similarité structurale et l'activité. Les composés bioactifs ont été rassemblés à partir de jeux de données présentant une confiance élevée dans la base de données ChEMBL.[4] Pour déterminer l'état d'ionisation d'un composé, deux types de valeurs calculées de constantes de dissociations (pKa) étaient disponibles dans ChEMBL. Les composés ont été classifiés comme appartenant à l'une des quatre classes (IS-class) : acide, basique, neutre et zwitterionique. Pour évaluer les relations structurales entre les composés de différentes IS-classes, les paires de molécules correspondantes ont été calculées (matched molecular pairs - MMPs). Une MMP est définie comme une paire de composés qui se distinguent uniquement par l'échange d'une sous-structure, souvent nommé "transformation chimique". Le voisinage chimique de chaque composé classifié a été déterminé en prenant en considération tous ses partenaires de MMP dans un jeu de données cible. Les états d'ionisation présentent une distribution inégale parmi les composés bioactifs. En majorité, les composés bioactifs étaient basiques ou neutres dans les conditions physiologiques. De plus, les ensembles à cible unique et les superfamilles cibles présentent des différentes significatives d'états d'ionisation. En outre, une analyse MMP systématique a révélé que des changements d'état d'ionisation apparaissent fréquemment parmi les analogues structuraux. Des analogues très actifs et très peu actifs diffèrent souvent par rapport à leurs IS-classes. Une préférence notable envers une IS-classe spécifique a été observée dans les composés hautement actifs pour différentes cibles, ce qui nous a offert un aperçu utile pour la conception et l'optimisation de composés.

Suivi de la progression relation structure-activité dans l'optimisation des têtes de série

L'optimisation de têtes de séries (Lead Optimization - LO) vise à transformer les composés actifs sélectionnés en candidats pour la recherche clinique au travers d'une évaluation itérative d'analogues. Le procédé est multiparamétrique par nature et est largement dirigé par des hypothèses et des règles empiriques variables, dépendant de l'expérience et de l'intuition des chimistes médicinaux. De plus, les projets de LO se déroulent souvent sur une longue période, nécessitant des ressources substantielles et générant des milliers de composés. Néanmoins, il n'est pas rare qu'un projet LO fasse face à des blocages et soit en fin de compte arrêté. Il est donc hautement désirable d'avoir à disposition un outil de diagnostic ou une métrique pour guider les chimistes médicinaux dans la progression du projet LO d'une façon objective. Nous introduisons ci-après un cadre de calcul basé sur les matrices de relations structure-activité (Structure-Activity Relationship Matrix - SARM) pour surveiller la progression de l'information sur les relations structure-activité (Structure-Activity Relationships - SAR) au cours du temps. Des jeux de données évolutifs ont été assemblés à partir de la ChEMBL et de deux projets différents de découverte de médicaments, chez Pfizer. Les SARMs ont été générées pour chacun de ces jeux de données à différents moments dans le temps: SARM "existante", "étendue" et "nouvelle". Ces SARMs ont été annotées à l'aide de deux scores numériques : un score de discontinuité basé sur la SARM ($\mathrm{SARM}_{\mathrm{Disc}}$) et l'activité médiane de tous les composés dans une SARM donnée. Les scores ont été visualisés dans des graphiques de dispersion afin de surveiller la distribution graphique des SARMs au cours du temps. Nous avons découvert que les décalages caractéristiques des ensembles SARMs dans les graphiques de dispersion indiquent une progression SAR positive, neutre ou négative. Au cours d'un projet de LO, une progression SAR est généralement observée quand des analogues sensibles au SAR et de plus en plus actifs sont créés. En prenant toutes ces informations en compte, nous démontrons que les indicateurs SARM peuvent être utilisés comme un outil de diagnostic pour surveiller la progression SAR au cours du temps. Cette approche peut être utilisée pour distinguer la progression

SAR de la redondance. Ces idées sont précieuses et devraient pouvoir appuyer le processus de prise de décision dans des projets de LO.

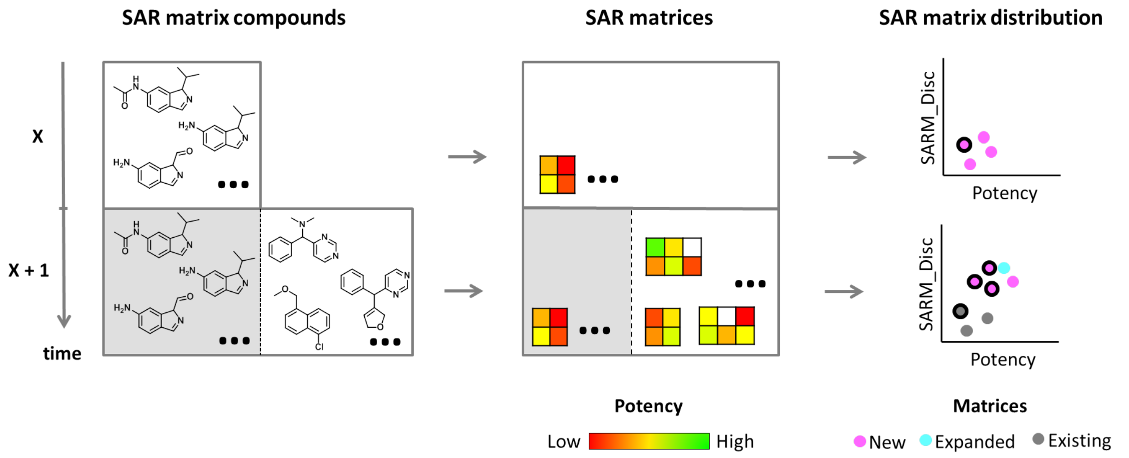


Figure 1: Représentation schématique illustrant le concept de surveillance de la progression SAR au cours du temps en utilisant les SARMs.

Détection et analyse de motifs structuraux privilégiés à l'aide des Generative Topographic Maps (GTMs)

Le terme de sous-structure privilégiée se réfère aux structures de base récurrentes dans les composés actifs contre une famille de cibles donnée, associées avec leur activité biologique. La recherche de sous-structures privilégiées est d'une importance particulière lors de la conception de nouveaux composés bioactifs. Dans cette étude, nous démontrons l'application de Generative Topographic Maps[5] (GTMs) pour extraire des motifs structuraux privilégiés à travers trois importantes superfamilles cibles (les protéases, les kinases, et les récepteurs couplés à la protéine G). Une GTM "universelle" précédemment publiée,[6] représentant une carte de l'espace chimique des composés drug-like, a été utilisée dans notre analyse. Un paysage GTM est composé de nœuds sur lesquels les composés sont projetés, sur la base des valeurs réelles d'association du composé à chaque nœud (responsabilités). Chaque composé sera caractérisé par son vecteur de responsabilité. Pour identifier des sous-ensembles de composés présentant des caractéristiques structurales communes, les vecteurs de responsabilité ont été convertis en une forme discrète appelée motif de responsabilité (responsibility pattern – RP). Les RPs ont été priorisés comme "privilégiés" (PRP) par rapport à une superfamille cibles, s'ils caractérisaient une grande fraction de composés appartenant à cette superfamille cibles mais que leur taux d'occurrence global (dans ChEMBL) était significativement plus bas. Les composés spécifiques de chaque PRP ont ensuite été soumis à une fragmentation basée sur des règles retrosynthétiques (RECAP), pour extraire leur "noyau-RECAP". Une analyse visuelle des noyaux-RECAP fréquents a permis d'identifier les "motifs structuraux privilégiés". Ceci fournit une façon alternative d'extraire des sous-structures privilégiées et étend l'approche classique centrée sur le motif structural.

Visualisation des relations structure-activité dans des larges jeux de données de composés

En chimie médicinale, la comparaison des caractéristiques structurales des composés à leur activité devient de plus en plus difficile lorsque la taille du jeu de données va au-delà d'une simple série d'analogues. Alors, la structure et l'activité doivent être comparées à travers un groupe de composés large et hétérogène. Pour ces cas, nous proposons un outil de visualisation à deux composantes, combinant une variante des generative topographic maps (GTM) basée sur des paires de composés avec un réseau de l'espace chimique (CSN), pour faciliter l'analyse des SAR dans des jeux de données grands et hétérogènes. L'outil GTM a été utilisé dans notre analyse pour apporter une vue "du satellite" globale des paysages d'activité. Celle-ci a été annotée, afin d'identifier des environnements locaux représentant des zones riches en information SAR. La couche suivante de notre outil de visualisation était la projection CSN des régions priorisées de la GTM. Le CSN a fourni une façon d'inspecter les communautés moléculaires contenant des informations sur les SAR, et a aidé à résoudre ces régions au niveau des composés individuels et de leurs relations. Notre schéma de visualisation GTM-CSN a été appliqué à des jeux de données d'optimisation (contenant de 1645 à 2202 composés) et à des jeux de données de criblage anti-malaria (contenant plus de 13000 composés). L'inspection visuelle des paysages d'activité de la GTM a révélé plusieurs régions d'intérêt, avec des jeux de données de criblage montrant des degrés de discontinuité SAR bas par rapport aux jeux de données d'optimisation. Les tendances SAR ont été explorées via les

CSNs, sur des îlots de SAR, pour analyser les modifications structurales résultant en de changements d'activité. Dans notre étude de preuve de concept, nous avons démontré avec succès que l'approche proposée de GTM-CSN facilite l'analyse SAR de grands jeux de données, et étend encore le spectre actuel des outils de visualisation de SAR disponibles.

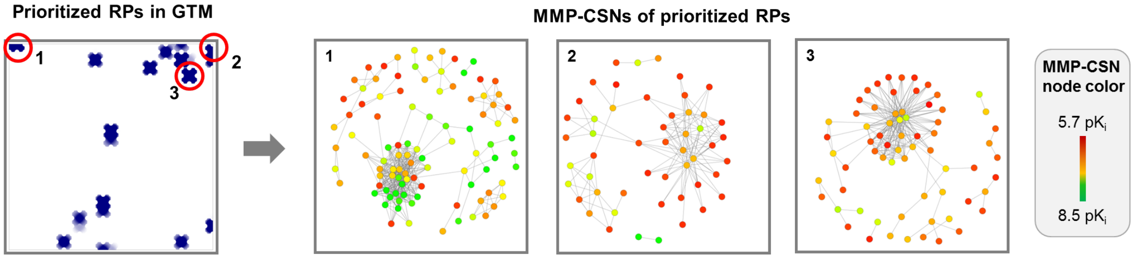


Figure 2: Les RPs prioritaires pour les jeux de données de criblage anti-malaria sont montrés sur une GTM simplifiée à gauche. Pour les sous-jeux de données de composés correspondants, les représentations CSN avec une disposition constante ont été générées. Les nœuds ont été colorés en fonction de l'activité des composés.

Visualisation de paysages à multipropriétés pour l'optimisation de composés

L'optimisation de composés repose sur des propriétés multiples et il est important, mais souvent compliqué, d'atteindre un équilibre entre elles. La méthode de visualisation par star coordinates (STC) est appliquée ici, pour aider à comparer des solutions numériquement équivalentes, et pour identifier les sous-espaces "drug-like" dans l'espace des propriétés multidimensionnel. Un espace de descripteurs à 14 dimensions, incluant l'activité des composés, a été conçu comme espace de référence pour l'optimisation multipropriétés. Un jeu de données de descripteurs chimiquement intuitifs a été calculé, incluant les caractères hydrophobiques et aromatiques, la complexité moléculaire, le potentiel de liaison hydrogène, la charge et les propriétés de surface. L'intérêt global d'un composé comme candidat bioactif peut être rendu par une fonction à multi-objectif (multi-objective function - MOF) de ses descripteurs. La STC, en tant que technique de visualisation multidimensionnelle, arrange les coordonnées (les descripteurs) dans des positions prédéfinies autour d'une origine commune au centre. La position de chaque composé dépend de la valeur de chaque descripteur. Quatre jeux de paramètres de poids pour chaque descripteur ot été calibrés, afin de trouver les combinaisons qui permet à de médicaments connus d'être associés de manière spécifique à des valeurs MOF elevées. Autrement dit, la valeur de MOF est calibrée pour correspondre à la “drug-likeness” d'un composé. Une inspection visuelle détaillée des STC aide l'identification des composés ayant des propriétés similaires à celles des médicaments. Les projections priorisées ayant des réglages de poids similaires pour leurs descripteurs et partageant la plupart des médicaments classés au premier rang ont présenté des distributions de composés distinctes dans l'espace de propriétés multidimensionnel. La visualisation STC est complémentaire à l'optimisation multidimensionnelle des propriétés des composés, et aide à différencier des solutions numériquement équivalentes avec des paramètres de propriétés similaires ou distincts.

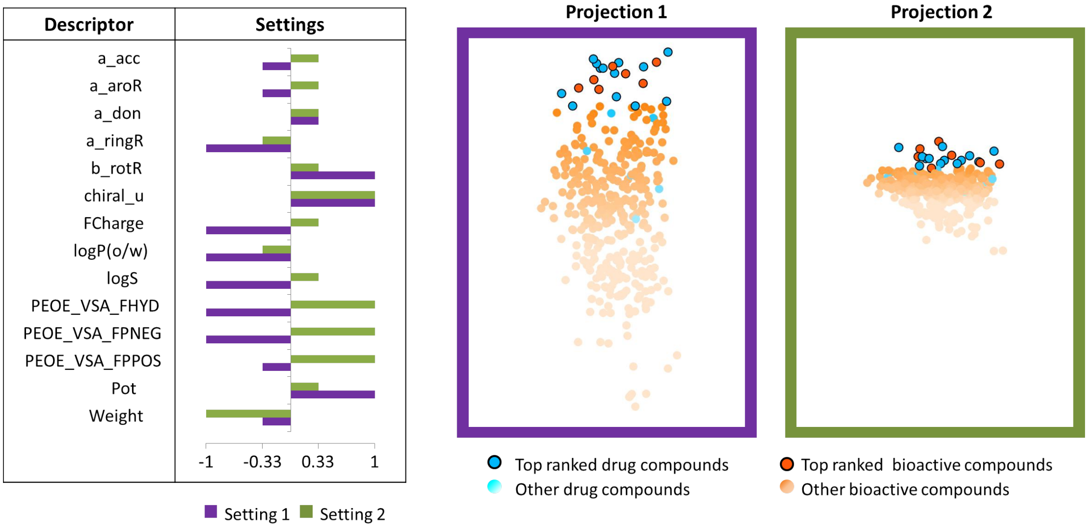


Figure 3: Deux paramètres de poids pour les descripteurs, et leurs projections de coordonnées star correspondantes sont visualisées et comparées, pour des ligands des récepteurs alpha-2a adrénergiques.

Prédiction de pics d'activités en utilisant les modèles de machines à vecteurs de support

Les pics d'activité (activity cliffs[7] - ACs) sont formés par des paires ou des groupes de composés actifs, qui sont structuralement similaires mais présentent de grandes différences d'activité. Ils dévoilent de petites modifications chimiques qui amènent à des effets biologiques profonds, et sont par conséquent d'un très grand intérêt pour la chimie médicinale. Deux nouvelles approches pour la prédiction par calcul des AC ont été basées sur : (i) les graphes de réactions condensés (condensed graphs of reactions - CGRs) et (ii) la recombinaison de descripteurs moléculaires (plain descriptor recombination - MQSPR). Le critère de similarité structurale requis pour les partenaires AC est l'existence d'un MMP constitué par ces molécules (qui se doivent donc être d'analogues proches différant uniquement par rapport à un seul substituant d'une – plus large – sous-structure commune). Un CGR représente un MMP comme une pseudomolécule, en encodant les partenaires de MMP dans un seul graphe moléculaire. Dans les MQSPRs, les descripteurs, à savoir le nombre de fragments pour chaque composé du MMP, ont été concaténés dans une empreinte numérique commune. Les paires de composés, encodés par les CGRs et les MQSPRs ont été utilisées comme données d'entrée pour des modèles de classification à vecteurs de support SVC et pour des modèles de régression à vecteurs de support SVR. En plus de la validation croisée classique à p-fold (p=3) "item-out", deux schémas de validation croisée plus avancés et plus stricts ("compound-out" et "both-out") ont été explorées. Dans le schéma "item-out", le tirage du tiers de tous les MMPs utilisé pour la validation est fait sans biais. Dans "compound-out", on denomine par tirage aléatoire un sous-ensemble de composés, et on garde toutes les MMPs auxquels ceux-ci participent dans le test de validation (en s'assurant que ces composés ne seront pas "vus" pendant le calibrage). Avec "both-out", aucun des composés faisant partie du jeu de données d'entrainement des MMPs n'a été utilisé durant l'étape d'entrainement. Les modèles SVC ont régulièrement atteint une haute performance. Dans le cas de la régression, les modèles basés sur la MQSPR ont surpassé les modèles basés sur les CGRs. De plus, une baisse de performance significative a été observée lorsque les schémas de validation "compound-out" et "both-out" ont été appliqués. Une classification précise des ACs a été atteinte en utilisant les modèles SVM, sans avoir besoin d'employer des noyaux sophistiqués. De plus, la régression SVM a été appliquée avec succès, pour la première fois, pour prédire la directionnalité des changements d'activité encodés par les MMPs.

Conclusions

Des méthodes computationnelles variées ont été présentées dans cette thèse, avec une démonstration claire de leur application dans l'analyse de l'espace chimique. Nos analyses ont exploité une abondance de composés structuralement distincts et hautement actifs, déjà disponibles pour de nombreuses protéines pertinentes en tant que cibles pharmaceutiques, en se focalisant sur des études de cas d'une relevance maximale (à partir des aspects liés aux super-familles de cibles, à des cibles spécifiques, à des propriétés systémiques - antipaludiques). En plus de la visualisation des paysages d'activité de grands sets de données, les GTMs ont été appliquées ici pour extraire des motifs structuraux privilégiés à partir de composés appartenant à différentes superfamilles cibles. Cette méthode de détection alternative pour détecter les sous-structures privilégiées est mise en œuvre sans connaissance préalable de la classification préalable des composés sur la GTM et s'étend au-delà de l'approche classique centrée sur le motif structural. Une prédiction de la différence d'activité signée dans les pics d'activité a été réalisée pour la première fois en utilisant des modèles SVM. En outre, des outils discutés ici tels que les SARMs et la STC offrent des conseils utiles aux chimistes médicinaux dans le cadre des projets d'optimisation de composés. Ensemble, les méthodes conçues ici représentent une panoplie assez complète d'outils de drug design, en allant de l'analyse globale "à vol d'oiseau" par cartographie, à l'analyse des discontinuités locales, à l'optimisation multimodale.

Communications scientifiques

- **S. Kayastha**, D. Horvath, G.Marcou, A. Varnek, J, Bajorath, "Prediction of Activity Cliffs using Condensed Graph of Reaction Representation, Descriptor Recombination , Support Vector Machine Classification and Regression", B-IT FOUNDATION COUNCIL AND ADVISORY BOARD MEETING, 15/11/2016 (poster)

- **S. Kayastha**, A. de la Vega de León, D. Dimova, J. Bajorath, "Visualization of Multi-property Landscapes for Compound Selection and Optimization", STRASBOOURG SUMMER SCHOOL IN CHEMOINFORMATICS 2016 (poster)

- **S. Kayastha**, “Analysis and Prediction of Activity Cliffs for Pharmaceutical Research”, SCIENTIFIC DAY OF UMR 7140, Strasbourg, 6/01/2017 (comm.orale)

- **S. Kayastha**, “Support Vector Classification and Regression Models for Activity Cliffs Prediction”, GUEST LECTURE FOR MASTERS IN CHEMOINFORMATICS, Strasbourg 18/11/2016

Publications dans le cadre de la these

- **Kayastha, S.**; Horvath, D.; Gilberg, E.; Gütschow, M.; Bajorath, J.; Varnek A. Privileged structural motif detection and analysis using generative topographic maps. *Journal of Chemical Information and Modeling* **2017**, *57(5)*, 1218-1232.

- Horvath, D.; Marcou, G.; Varnek, A.; **Kayastha, S.**; de la Vega de León, A.; Bajorath, J. Prediction of activity cliffs using condensed graphs of reaction representations, descriptor recombination, support vector machine classification, and support vector regression. *Journal of Chemical Information and Modeling* **2016**, *56*, 1631-1640.

- Shanmugasundaram, V.; Zhang, L.; **Kayastha, S.**; de la Vega de León, A.; Dimova, D.; Bajorath, J. Monitoring the progression of structure-activity relationship information during lead optimization. *Journal of Medicinal Chemistry* **2016**, *5*, 4235-4244.

- de la Vega de León, A.; **Kayastha, S.**; Dimova, D.; Schultz, T.; Bajorath, J. Visualization of multi-property landscapes for compound selection and optimization. *Journal of Computer Aided Molecular Design* **2015**, *29*, 695-705.

- **Kayastha, S.**; de la Vega de León, A.; Dimova, D.; Bajorath, J. Target-based analysis of ionization states of bioactive compounds. *Medicinal Chemistry Communications* **2015**, *6*, 1030-1035.

- **Kayastha, S.**; Dimova, D.; Stumpfe, D.; Bajorath, J. Structural diversity and potency range distribution of scaffolds from compounds active against current pharmaceutical targets. *Future Medicinal Chemistry* **2015**, *7*, 111-122.

Chapter 1

Introduction

1.1 Chemoinformatics in Drug Discovery

Chemoinformatics is an interdisciplinary field whereby computational methods are applied to tackle chemical problems. The term “chemoinformatics” was first introduced in literature by Brown in 1998 as “the mixing of all the information resources to transform data into information and information into knowledge for intended purpose of making better decisions faster in the arena of drug lead identification and optimization”.[8] Despite being introduced as a relatively new term, many chemoinformatics approaches were initiated in 1960s and early 1970s.[9] Moreover, the term chemical informatics was used much earlier to describe the application of informatics approaches to chemistry, albeit lacking specific focus on drug discovery.[10] Later Hann and Green pointed out that chemoinformatics is indeed a new name for an old problem.[11] At present, chemoinformatics covers a broad spectrum of computational methodologies including, compound database mining, library design and optimization, molecular similarity and diversity analysis, chemical structure and property prediction to name a few.[12] The studies presented in this thesis mainly focus on the development and application of chemoinformatic approaches to address challenges faced in modern drug discovery.

Drug discovery is a process of discovering and developing novel pharmaceutical agents to treat a specific disease condition. It is a complex procedure involving multiple stages and is carried out in the following sequence:

1. **Target selection**: Drug discovery projects start with identification of a suitable target, to which a drug molecule is able to bind and elicit a biological response. Biological targets can range from proteins, genes to RNAs. Once the target is identified, it is validated using *in vitro* techniques or disease models.

2. **Hit identification**: During this phase, compound screening assays are utilized to identify hits, i.e., compounds that show specific biological activity against the target of interest and whose activity is confirmed upon re-testing.[13] A variety of screening strategies exist which involve different assay of compound sets in high-throughput fashion.[14]

3. **Lead optimization**: Once hit series have been identified, each of them are further refined via synthetic exploration and subsequent testing to generate more potent and selective compounds called leads. In addition, the absorption, distribution, metabolism, excretion, toxicity (ADMET) and other physicochemical properties of lead molecules are tested to ensure a good balance between their affinity, selectivity and metabolic properties.[14]

4. **Pre-clinical development**: Compounds prioritized in lead optimization stage undergo *in vitro* or *in vivo* tests during pre-clinical development, to determine if they are suitable and safe to be administered to humans in the next stage.

5. **Clinical development**: In this final stage of drug discovery, the candidate compound that passed all of the previous phases are tested in human volunteers via clinical trials. Experimental studies are carried out to explore the dose and exposure levels tolerable by humans.[15] Additionally, any adverse effects that may arise through the use of the candidate compound is closely monitored in patients.

After sufficient proof highlighting the safety and efficacy of the candidate compound has been gathered, it should be reviewed by appropriate regulatory body before getting approval for marketing it as a new drug.

Drug discovery projects typically last for periods of 10-15 years costing over a billion dollars.[13] Development of a drug from the initial stage of target selection through lead identification and optimization to making it available in the market is thus a highly resource and time consuming process. Modern drug discovery process has been making efforts to minimize the time and costs of research and development stages by incorporating a variety of techniques from chemoinformatics.[16] One of the major aims of chemoinformatics approaches focusing on drug discovery is to understand the interaction between small molecules and their biological targets. By uncovering the structural features that determine the biological activities of small molecules, rational decisions can be made during lead optimization stage ultimately paving a way to identification of novel compounds with desired biological activities.

It should be noted that biological activity is only one of the many properties of a compound that should be optimized to qualify as a successful drug candidate. Given the multi-parametric nature of compound optimization, computational multi-property optimization approaches have been used, attempting to reach a reasonably good balance between drug-relevant properties of a compound. High throughput screening techniques have generated huge amount of compound activity data, expanding the current universe of bioactive compounds. Prioritizing sub-spaces representing compounds with desired activities/properties in such a huge chemical space is a rather daunting task. Data mining and visualization approaches have therefore proven helpful in navigating through chemical spaces to explore and ultimately identify promising drug candidates.

1.2 Molecular Representations

Molecular representations provide a way to encode structural features and properties associated with a given molecule in an interpretable way. Molecular representations vary in their complexity depending on the way they are derived and the amount of information they capture. The simplest molecular representation is one-dimensional (1D), such as the molecular formula. It can be used for determining the chemical composition of a given molecule but this is not very informative. Molecular graphs are two-dimensional (2D) representations of chemical structures, such that nodes and edges correspond to atoms and bonds, respectively. The topology of a given molecule, i.e., the connectivity of atoms and bonds, is clearly indicated in molecular graphs. This information is translated into a computer-interpretable form by connection table or linear notations. In addition to molecular topology, connection tables contain atom coordinates, their hybridization states and bond orders. Linear notations like simplified molecular input line entry specification (SMILES)[17] and IUPAC international chemical identifier (InChI)[18] are strings of alphanumeric and special characters constructed following predefined rules for representing molecular structures. Due to their compact form, they have become more popular in recent years, mainly for representing molecules in large databases. 2D molecular representations are not capable of describing the spatial positions of atoms and bonds, which determine molecule's steric and electronic properties. Three-dimensional (3D) representations such as molecular surface and volume capture molecule's key conformational properties based on the spacial arrangements of the atoms.

1.2.1 Scaffolds

Scaffold represents structural backbone of a given molecule. The scaffold concept is widely used in medicinal chemistry to describe core structures of bioactive compounds.

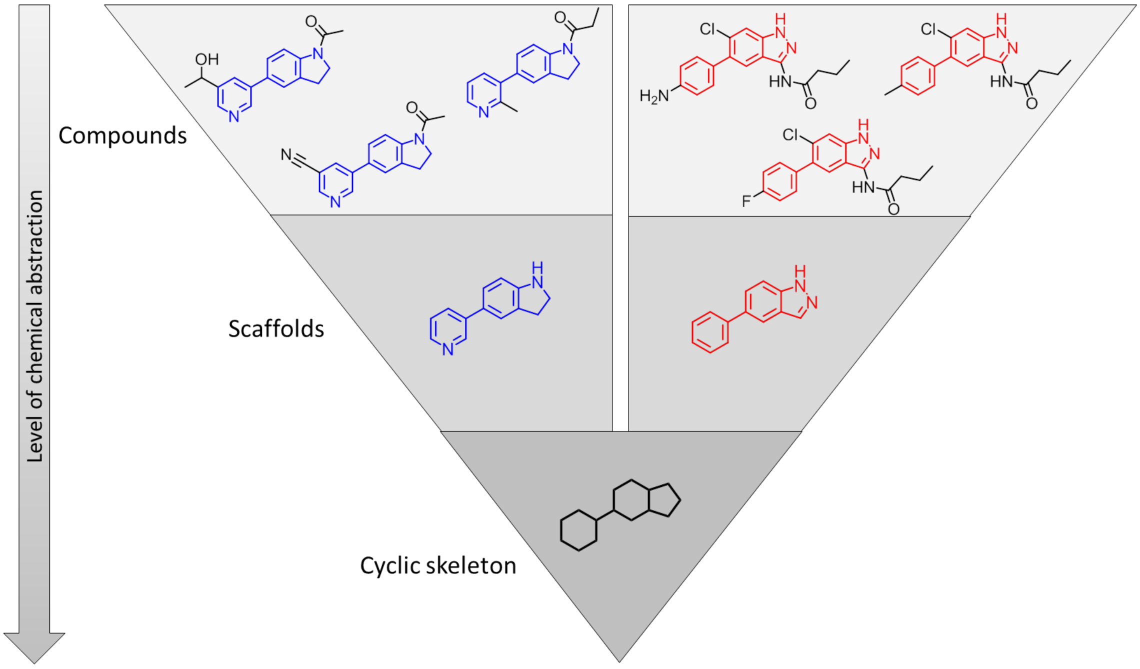


Figure 1.1: Scaffolds and cyclic skeletons. For two exemplary sets of compounds, corresponding scaffolds and CSKs are given. Compounds within a set represent structural analogs and have a common scaffold, highlighted in blue and red for the left and right compound sets, respectively. Further, both scaffolds give rise to a single CSK, thus indicating different levels of chemical abstraction, from compounds through scaffolds to CSKs.

It provides an important means of structural organization and classification of compounds in large compound databases or screening libraries.[19] Scaffolds are also linked to their biological activity, often to search for privileged substructures or identify scaffold hopping instances. Privileged substructures are defined as core structures associated with compounds that are preferentially active against a specific group of targets. Computational compound screening is largely motivated by scaffold hopping, i.e., identification of compounds with structurally distinct cores, but same or similar activity. Scaffold is a rather loosely defined term and many different definitions of scaffold have been put forward.[20] The most widely applied definition was given by Bemis and Murcko, in which scaffolds are obtained from compounds by removing all non-ring R-groups while retaining all ring structures and linker fragments connecting two or more ring structures.[3] A further chemical abstraction called cyclic skeletons (CSKs), are obtained from scaffolds by converting all hetero atoms to carbon and all bond orders to one (**Figure 1.3**).[21]

1.2.2 Matched Molecular Pairs

Matched molecular pairs (MMPs) are defined by a pair of compounds that only differ by structural modification at a single site.[22] MMP partners can hence be interconverted to one another by exchange of well-defined substructures, often termed as chemical transformations (**Figure 1.4**). As outlined earlier, molecular similarity is subjective and the outcome of similarity assessment heavily relies on the chosen molecular representation. Studies have shown that SAR information of compounds change when different fingerprint representations were used.[23] Moreover, numerical similarity metrics are not always chemically interpretable. MMPs have an intrinsic advantage over other molecular representations in this respect for more than one reason. To assess structural similarity between compound pairs, the MMP-based approach does not require any predefined structural relationship. In addition, chemical transformations can be limited to only small structural changes as seen in analog series by introducing well-defined size restriction criteria. In contrast to numerical similarity measures, MMP-based similarity is well-defined and chemically intuitive, therefore making it easier for medicinal chemists to identify SAR patterns. In recent years, MMP analysis has gained a lot of popularity and it has been used in wide range of computational medicinal chemistry applications, mainly large scale SAR analysis and visualization. MMP formalism has been applied for identification of activity cliffs (vide infra), here termed as "MMP-cliffs", by replacing calculated similarity metrics with well-defined substructure relationships.[24]

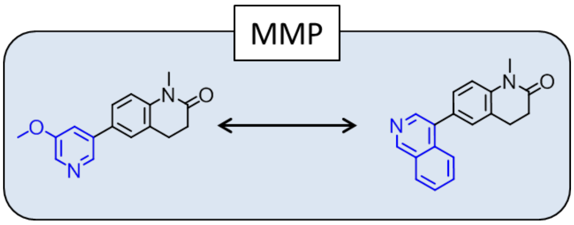


Figure 1.2: Matched molecular pairs. Two compounds participating in a matched molecular pair are shown. Substructures highlighted in blue represent chemical transformation and the remaining part of the molecules represents the core structure shared by both compounds.

Given the increasing interest in MMP analysis, a number of conceptually different algorithms have been introduced for identification of MMPs from compound data sets. One of the first algorithms for MMP generation was based on a predefined list of chemical transformations that was searched against all possible compound pairs in a given data set.[25–27] Despite low computational complexity of this method, MMPs that were identified were limited within those present in the predefined list. Maximum common substructure (MCS)-based algorithms overcome such limitations by systematically enumerating all MMPs for a given compound set. In this method, largest substructure shared by compound pairs are identified as common cores whereas the remaining part of the structures are regarded as the chemical transformations. Chemical modifications in MMP partners thus correspond to only small structural modifications as a large part of their structure are shared as MCS. MCS-based algorithms tend to be computationally expensive and are often combined with clustering techniques and subsequent pruning of the data set, which might limit its performance.[28,29]

A more recent approach follows a two-step procedure that requires fragmentation of molecules for MMP calculation. In the first step, data set compounds are decomposed into well-defined fragments. The second step involves indexing of generated fragments and subsequent comparison and identification of compound pairs with common substructures. Fragmentation-based algorithms can be applied to very large data sets and do not require huge computational resource like MCS-based approaches because each molecule is processed only once. A widely used fragmentation-based MMP algorithm was introduced by Hussain and Rea.[30] Molecules are first fragmented by systematically deleting all non-ring single bonds between two non-hydrogen atoms, creating so-called single-cut fragments. The larger of the two fragments is termed as "key" and the other one as "value". Combinations of two or three single bonds are also used for generating double- and triple-cut fragments. Once the fragments are generated, they are stored in an index table as key-value pairs. Entries of the index table in which a single key has multiple values represent MMPs, the key being the common substructure and value the chemical transformation. If a key has more than two value fragments, the corresponding compounds form a matched molecular series (MMS). The compounds in an MMS therefore must form all possible MMP relationships among each other.

1.2.3 Molecular Descriptors

Substructure-based molecular representations like scaffolds and MMPs are chemically intuitive and interpretable. Alternatively, molecules can be represented in terms of numerical values called descriptors. Molecular descriptors are mathematical functions or models that represent measured or predicted properties associated with molecules. A wide range of descriptors have been introduced,[31–33] and can be classified as 1D, 2D or 3D molecular descriptors. 1D descriptors are simple as atoms counts and molecular weight, which can be easily derived from molecular formula. 2D descriptors are mainly calculated from 2D molecular graphs such as topological index and molar refractivity. 3D molecular descriptors like 3D pharmacophore keys and molecular surface areas require 3D confirmation of molecule.

It should be noted that the choice of descriptors depends on the task in hand, since the importance of molecular features varies with respect to specific application. Therefore, the descriptor selection should be done on a case-by-case basis.

In our studies, 2D descriptors originating from the following two software platforms were used and are discussed in detail in the following.

- **In silico design and data analysis (ISIDA) fragmentor**:[34] It is a software developed as a part of ISIDA project for calculating molecular fragment count descriptors based on a series of graph algorithms described in the book Algorithmes de graphes.[35] There are three major classes of ISIDA descriptors developed so far:

  1. substructural molecular fragments (SMF)[36,37]
  2. fuzzy pharmacophoric triplets (FPT)[38,39]

3. property-labelled fragments(IPLF)[40]

We used SMF and IPLF fragments generated with ISIDA fragmentor 2015. Substructural molecular fragments enumerate the occurrence of fragments of differing type and length within a molecule. The fragments can be categorized into any of the following.

- **Sequences** of successive atoms or/and bonds of a predefined length representing shortest possible path between two given atoms.
- **Augmented atoms** encoding the connectivity information of a selected atom with its neighboring atoms or/and bonds to a given topological distance (sphere).
- **Triplets**, i.e. all possible combinations of three atoms in a molecular graph along with the topological distances between them. For example, the triplet formed by the atoms numbered 4, 7 and 12 of the molecular graph in **Figure 1.3** will yield a triplet N5C4O5, where distance between atoms 4 and 7 is d(4,7)=5. Similarly, d(7,12)=4 and d(4,12)=5.

Figure 1.3 shows the substructural molecular fragments obtained from an exemplary molecular graph.

IPLF descriptors also monitor counts of specific subgraphs of the molecule but the vertices of the molecular graph that are employed for their calculation are colored by properties of the atoms, mainly by pharmacophoric features and electrostatic potential. An example of molecular graph coloration based on pharmacophoric rules is given in **Figure 1.4**. Once atoms of the molecule are annotated by a given feature, the program calculates fragment counts under the desired fragmentation scheme (sequences, augmented atoms, etc).

- **Molecular Operating Environment (MOE)**:[41] MOE is a software package that has integrated various tools facilitating drug discovery research including ligand- and structure-based drug design, molecular modeling and simulations, protein and antibody modeling, and pharmacophore discovery. QuaSAR descriptor module available in MOE package allows users to calculate more than four hundred 2D and 3D molecular descriptors that can be used for QSAR modeling, similarity and diversity analysis or combinatorial library design.[42]

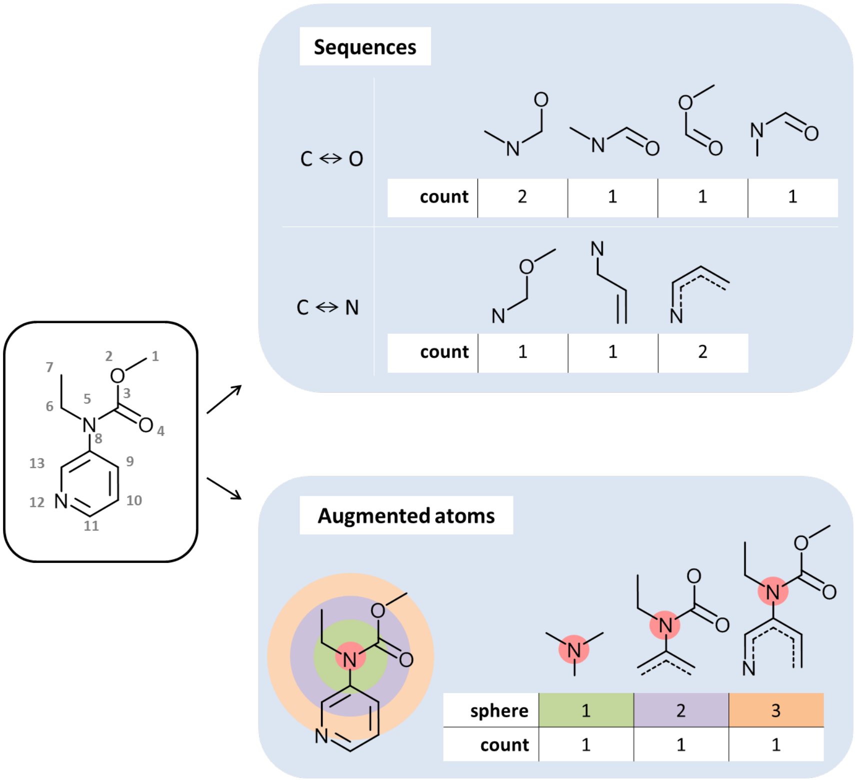


Figure 1.3: ISIDA substructural molecular fragments. Shown are the substructural molecular fragments and their counts generated from an exemplary molecule. Sequences of atoms and bonds of length four between two different atoms in the molecule (carbon and oxygen; carbon and nitrogen) are given. Augmented atom fragments calculated at topological distances of one (highlighted in green), two (purple) and three (orange) from the central nitrogen atom (red) are shown.

> 2D QuaSAR descriptors include atom and bond counts (number of heavy atoms, number of rotatable bonds, etc.), physico-chemical properties (molar refractivity, atomic prolarizabilities, LogP, etc.), connectivity and shape indices (atomic connectivity index, Kappa shape index, etc.) as well as several pharmacophore and partial charge descriptors. In addition, 3D descriptors accounting for potential energy, surface area, volume, shape, and conformation dependent charge can also be calculated using QuaSAR module. Furthermore, users have an option to define their own custom descriptors in MOE.

Molecular fingerprints, a subtype of molecular descriptors, are bit or integer string representations of molecular structure and properties. They encode different 2D and 3D molecular features in form of bit string representation indicating presence or absence of features or integer string recording their frequency of occurrence.

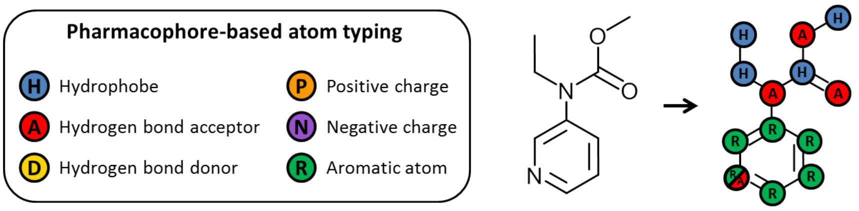


Figure 1.4: Pharmacophoric graph coloration. Given are the pharmacophoric rules for coloring atoms of an exemplary molecule prior to calculation of ISIDA property-labelled fragments.

Originally developed for substructure searching, a number of different fingerprint design have been introduced over the years that differ in their complexity, composition and length. One of the classical fingerprints that are still in wide use across many chemoinformatics applications are substructural fingerprints. Also known as “keyed fingerprints”, they are represented as a feature vector of fixed length, with each bit position corresponding to a predefined structural fragment. One prime example of substructural fingerprint is molecular access system (MACCS)[43] fingerprints, which contains 166 predefined structural keys.

Combinatorial fingerprints encode topological connectivity information of molecules and are generated by enumerating all possible paths through the molecule at a predefined length. The connectivity pathways are then mapped to overlapping (multiple) bit positions of the fingerprint of fixed length using a hashing function. In contrast to structural keys, hashed fingerprints do not have one-to-one correspondence of feature and bit positions, and thus are not chemically interpretable. Hashed fingerprints from Daylight Chemical Information Systems Inc.[44] contain 1024 bits, out of which typical organic molecules set between 50 to 400 bits.[45] More recent variants of combinatorial fingerprints such as Molprint 2D[46,47] and extended connectivity fingerprints (ECFP), capture layered atom environment of a given molecule up to a predefined bond diameter. ECFP4 accounts for the connectivity of atoms around each non-hydrogen atom of a molecule up to a four-bond diameter to define atom neighborhoods of increasing size. The length of these fingerprints is not fixed as atom environment calculations are molecule-specific.

Pharmacophore fingerprints are derived from 2D molecular graph or 3D conformation of molecules and several variations of these fingerprints exist. Pharmacophores are defined are parts of a molecule that are essential for its binding to the receptor and are important determinants of their biological activity. Typical examples include, hydrogen bond donors and acceptors, hydrophobic centroids, aromatic rings, etc. Pharmacophore fingerprints are based on two-, three-, or four-point pharmacophore models that capture such pharmacophore features and their inter-feature distances.

1.3 Molecular Similarity

Molecular similarity provides a way of grouping compounds based on their structural features, biological activity or physicochemical properties. It is a very popular concept in chemoinformatics and drug discovery. Over the years, a steady growth has been observed in a number of studies related to molecular similarity.[46] It plays a key role in lead discovery and compound optimization. Characterization of SAR features requires similarity assessment measures in order to relate structure and biological activity of compounds.

Evaluation of molecular similarity is subject to individual perspective. Chemoinformatic approach towards molecular similarity compares features of the molecules globally, whereby quantitative readout of structural relatedness are used for applications such as large-scale structure activity relationship analysis (vide infra) and similarity searching (vide infra).[48,49] On the other hand, a medicinal chemist's perspective of molecular similarity focuses mainly on local features, such as specific structural patterns that are important for biological activity. In addition, similarity assessment is governed by intuition, knowledge and experience of medicinal chemists. Equally experienced medicinal chemists may arrive at different conclusions in prioritizing preferred compounds based on their features.[50–52] Moreover, such similarity comparison is heavily context-dependent and resulting analysis may differ with respect to the order in which the medicinal chemistry expert views the compounds.

Computational similarity assessment largely depends on the way in which molecules are represented and the similarity metric used for calculation.[53] Moreover, the degree of similarity is subject to structural diversity of compound class being studied.[53] Molecular similarity assessment is a two-step procedure that involves first, choosing a molecular representation and secondly, applying a method to to extract information encoded in such representations and determine similarity between compounds. If molecules are represented in form of descriptor vectors then a suitable similarity metric is used. In case of substructure-based molecular representations, structural rules are defined to compare molecules. There is no standard or universally accepted method to represent molecular and chemical information, and thus the choice of representation is task dependent.

1.3.1 Similarity Metrics

Molecular similarity between compounds represented as bit strings can be quantified by similarity metrics. A number of similarity metrics have been introduced thus far, such as Tanimoto coefficient (Tc), Dice coefficient (Dc), Tversky coefficient (Tv) and Cosine coefficient.[53,54] Among them, Tanimoto coefficient is the most widely used similarity metric in chemoinformatics. If a and b represents the number of features present in compounds A and B, respectively, and c is the number of features shared by both, Tanimoto coefficient between these two compounds is given by:

$$Tc(A,B) = \frac{c}{a+b+c} \tag{1.1}$$

If two compounds do not have a single common feature, then Tc=0. As the number of common features increases, the Tc value also increases. If two feature sets are identical, then Tc=1. Therefore, the value of Tc similarity between two compounds always falls in the interval [0,1]. However, one should keep in mind that Tc=1 does not imply identical structures but only the presence of identical features between two compounds.

Instead of similarity, sometimes dissimilarity is calculated as the complement of similarity (i.e., dissimilarity = 1 - similarity). Smaller values of dissimilarity coefficient imply that two compounds are similar and vice versa. Euclidean distance, Hamming distance and Soergel distance are some of the examples of dissimilarity metrics.[53,54]

1.4 Structure-Activity Relationships

Structure-activity relationship (SAR) analysis represents a central theme in medicinal chemistry research, particularly for compound optimization efforts. SAR aims to establish relationship between structural features of compounds and their biological activity or potency. Traditionally, exploration of SAR was done in a case-by-case basis focusing on individual compound series and was highly subjective to chemical intuition and experience of the medicinal chemists. Conventional R-group tables were widely used as a qualitative means to record SAR patterns in analog series. An R-group table assembles core and substituent information of active compounds along with their potency in a tabular format to identify functional groups that might determine specific biological activities. Other qualitative SAR analysis methods, like similarity searching, involves assessment of whole-molecule similarity. In similarity searching, known active compounds are used as queries to search and rank the database compounds in decreasing order of their similarity to the query molecules. The top ranking molecules are thus expected to have biological activity similar to that of the query molecule.

Unlike similarity searching where all features are compared between query and database compounds, quantitative SAR (QSAR) methods focus only on key features of compounds that determine their biological activity. 2D and 3D QSAR studies have been carried out for more than four decades for computational SAR modeling and analog design.[55,56] The major goal of QSAR modeling is to establish a relationship between compound structures and their activity for a set of compounds with known activity values. Models with good predictive performance are then applied to predict potency of newly designed analogs. A major drawback of QSAR approach therefore is that its applicability domain is limited to structural analysis of only the immediate chemical neighborhood of the compound series which was used to derive the relationship. It is often difficult to achieve reliable predictions for test compounds that differ from the chemotypes used in model generation.

Clearly, both QSAR and similarity searching methods rely on the fundamental aspect of similarity property principle (SPP), which states that "similar molecules should have similar biological activities".[57] Therefore, it is of utmost importance that the molecular representation should be chosen such that higher structural similarity between compounds also reflect similar biological activity. The applicability domain of QSAR and similarity searching methods is thus limited to SAR continuity,[58,59] i.e. gradual structural changes leading to gradual changes in biological activity. However, SAR landscape of a given data set may comprise of other SAR characteristics that are not complaint with SPP intrinsic to QSAR and similarity search tools. SAR discontinuity is one such phenomenon and it refers to the occurrence of small structural changes causing significant effects in biological activity of compounds. Large scale SAR analysis methods are designed to explore various SAR characteristics of a given data set, including SAR continuity and discontinuity. Further, these methods are able to distinguish between different local SAR environments that give rise to global SAR heterogeneity.[60] Therefore, large-scale SAR exploration are descriptive rather than predictive in nature, mainly focusing on data mining and information extraction from large data sets of several hundred to thousands of compounds.

Structure-activity similarity (SAS) map[61] captures SAR characteristics of a given data set in form of a 2D scatter plot. Each data point of SAS map corresponds to a compound pair and axes represent structural similarity and potency difference values between them. A schematic representation of an SAS map is given in **Figure 1.5**.

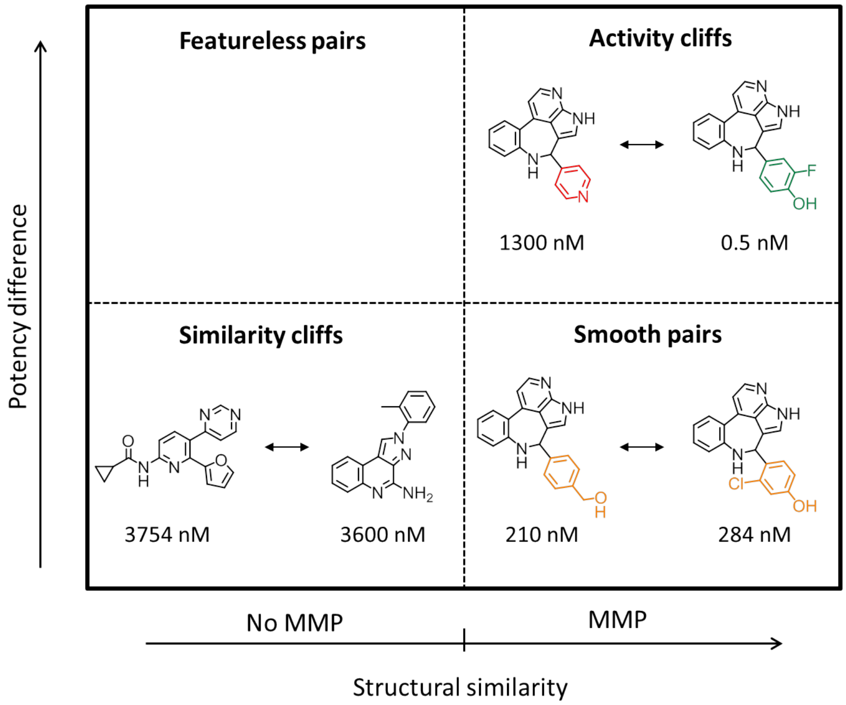


Figure 1.5: Strutcure-activity similarity maps. A schematic representation of SAS map is given, highlighting regions characterized by different SAR features. Each data point mapped onto an SAS map represents a compound pair and x- and y- axis corresponds to their structural similarity and potency difference measurements, respectively. Structural similarity criterion is defined in terms of MMP. Exemplary compound pairs involved in each of the SAR features and their potency measurements are shown. Activity cliff and smooth pairs are formed by structurally similar (MMPs) compounds and the corresponding chemical transformations are highlighted in red, yellow or green for weakly, moderately or highly potent compounds, respectively. Similarity cliff forming compounds do not participate in MMP but have comparable potency values.

Depending on predefined similarity and potency difference thresholds, regions of the map with varying information content can be identified. “Activity cliffs” shown at the upper right corner of the SAS map represent the extreme form of SAR discontinuity. Activity cliffs are defined by structurally similar compound pairs with significant difference in their potency,[7] and hence provide an immediate access to small structural changes that translate to significant differences in activity. The lower right quadrant is occupied by "smooth pair" compounds, representative of SAR continuity and are essential for QSAR modeling and prediction. The lower left regions of SAS maps correspond to "similarity cliffs", i.e., pairs of compounds with high activity similarity but low structural similarity. Highly potent compound pairs of these regions are of interest, mainly in search for structurally distinct compounds with similar activity. "Featureless pairs", formed in the upper left region are less interesting and represent compound pairs with low activity and structural similarity.

1.4.1 Machine Learning for SAR Analysis

Machine learning methods aim to develop computational models by learning from the data to derive rules and make predictions. Machine learning algorithms such as neural networks,[62–64] decision trees,[65] random forest,[66] etc., have become increasingly popular for developing SAR models. Apart from predicting novel active molecules, these machine learning models have been applied for predicting physicochemical properties, compound classification, and ligand-based virtual screening.

Support vector machines (SVMs)[67] are one of the most widely used machine learning methods in drug discovery and are able to perform comparably or even better than other state-of-the-art machine learning techniques.[68,69] Support vector machines are supervised learning[70] methods, originally introduced for binary classification of instances.[71] The general idea is to introduce a maximum-margin separating hyperplane in high-dimensional reference space in order to distinguish objects belonging to different classes.

SVM Classification

Learning phase of SVM utilizes a set of training instances $x_{\text{i}}, y_{\text{i}}, (i = 1, ..., n)$ with $x_{\text{i}} \in R^{\text{d}}$ being a feature vector and $y_{\text{i}} \in \{-1, 1\}$ the class label (positive or negative) of a training object i. A hyperplane H that best separates positive and negative instances is defined by the normal vector w and bias b as follows:

$$H = \{x \,|\, \langle x_{\text{i}}, w \rangle + b = 0\} \tag{1.2}$$

where, $\langle \cdot, \cdot \rangle$ is a scalar product.

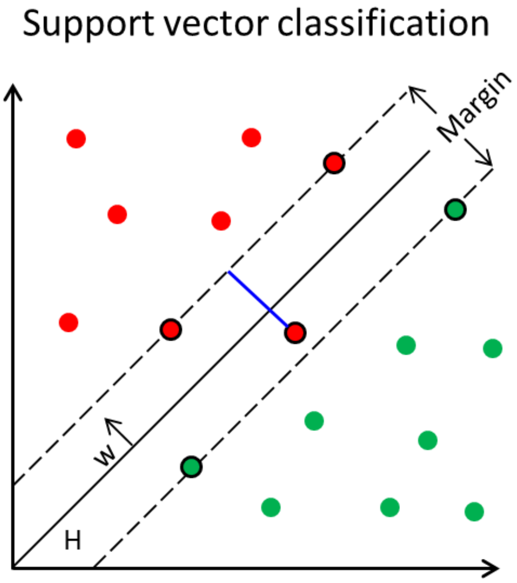


Figure 1.6: SVM classification SVM classification is shown for separating two classes indicated by red and green dots, respectively. The separating hyperplane (H) is shown as a solid black line and two dotted lines parallel to it represents the margins. Support vectors (dots encircled in black) lies either on the margin, within two margins or on the incorrect side of the separating hyperplane. Incorrectly classified data points are assigned higher values of slack variables (blue line).

An infinite number of hyperplanes can be obtained for correct classification of a linearly separable training data. Out of them, SVM chooses the one that maximizes the so called "margin", i.e., the distance between the hyperplane and the nearest training instances. To ensure the correct classification of all training instances by the optimal hyperplane, following condition must be satisfied:

$$y_{\mathrm{i}}(\langle x_{\mathrm{i}}, w \rangle + b) \geq 1 \quad \forall\, i \tag{1.3}$$

The training examples (both positive and negative) that are closest to the hyperplane H is known as "support vectors". The distance between the "support vectors" and the hyperplane is given by $\frac{1}{||w||}$. An optimal hyperplane thus maximizes $\frac{1}{||w||}$, or conversely, minimizes $||w||$.

When the training data is noisy and not linearly separable, a direct solution to the minimization problem cannot be obtained. In such cases, the constraints in **Equation 1.3** are relaxed by introducing non-negative "slack variables" $\xi_{\mathrm{i}}, i = 1, ..., n$.

The slack variables allow some of the support vectors to lie within the "margin" or even move across the hyperplane towards the incorrect side. The larger the value of slack variables, the higher is the misclassification of the training examples. In order to penalize the misclassification of training data, a parameter $C > 0$ is introduced. The resulting minimization problem is given as follows:

$$\text{minimize: } \frac{1}{2}||w||^2 + C\sum_{i} \xi_{\mathrm{i}} \tag{1.4}$$

$$\text{subject to: } y_{\mathrm{i}}(\langle x_{\mathrm{i}}, w\rangle + b) \geq 1 - \xi_{\mathrm{i}} \quad \text{with} \quad \xi_{\mathrm{i}} \geq 0 \quad \forall\, i \tag{1.5}$$

The optimization problem given in **Equations 1.4** and **1.5** can be reformulated using Lagrange multipliers,[72] α_{i}, as given below:

$$\text{minimize: } L_{\mathrm{D}} = \sum_{i} \alpha_{\mathrm{i}} - \frac{1}{2}\sum_{i,j} \alpha_{\mathrm{i}}\alpha_{\mathrm{j}}y_{\mathrm{i}} \langle x_{\mathrm{i}}, x_{\mathrm{j}}\rangle \tag{1.6}$$

$$\text{subject to: } \alpha_{\mathrm{i}}y_{\mathrm{i}} = 0 \quad \text{with} \quad 0 \leq \alpha_{\mathrm{i}} \leq C \quad \forall\, i \tag{1.7}$$

The solution to above convex optimization problem is defined by an α that uniquely determines the normal vector of the hyperplane as $w = \sum_{\mathrm{i}}\alpha_{\mathrm{i}}y_{\mathrm{i}}x_{\mathrm{i}}$. The "support vectors" are assigned a non-zero α_{i} and thus are the sole determinants of the position of the hyperplane.

The final decision function to classify a given test object x is formulated as follows:

$$f(x) = sgn\left(\sum_{i} \alpha_{\mathrm{i}}y_{\mathrm{i}} \langle x_{\mathrm{i}}, x\rangle + b\right) \tag{1.8}$$

The test set example x is classified into positive or negative class, depending on the values 1 or -1 as returned by the decision function, respectively. **Figure 1.6** shows the schematic representation of SVM classification.

SVM Regression

SVM regression, also known as support vector regression (SVR) is used for regression analysis to predict real values of the objects rather than their class labels.[71] The training data is defined by $\{x_{\mathrm{i}}, y_{\mathrm{i}}\}(i = 1, 2, ..., n)$ where x_{i} is a vector from the input space X and y_{i}, a real number associated with it. SVR aims to derive a function $f(x) = \langle w, x \rangle + b$, such that the training data is closer to the output y_{i} with the deviation not exceeding ξ.[73] The convex optimization problem is given by:

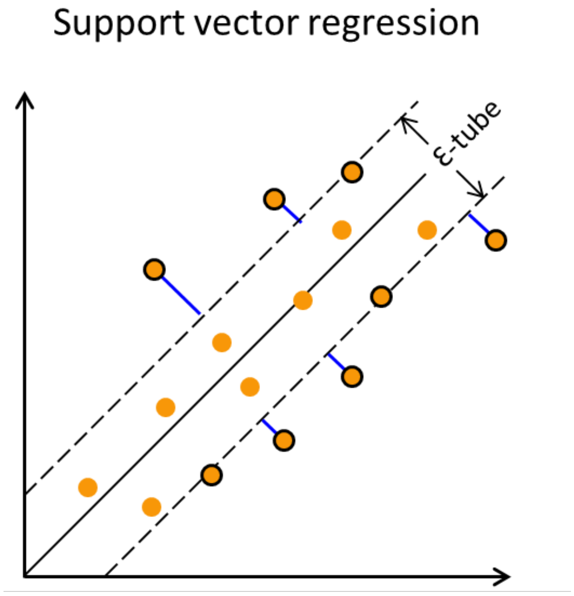


Figure 1.7: SVM regression Support vector regression is shown. Regression line (solid black line) is fitted to the objects (orange dots). ϵ-tube is shown by the region enclosed by the two dotted parallel lines. Support vectors are encircled in black and lie either on the edge or outside of the ϵ-tube. Higher slack variables (blue line) are assigned to objects that go further away from the ϵ-tube.

$$\text{minimize: } \frac{1}{2}||w||^2 + C\sum_{i}\left(\xi_{\mathrm{i}} + {\xi^*}_{\mathrm{i}}\right) \tag{1.9}$$

$$\text{subject to: } \begin{cases} y_{\mathrm{i}} - \langle x_{\mathrm{i}}, w \rangle - b & \leq \epsilon + \xi_{\mathrm{i}} \\ \langle x_{\mathrm{i}}, w \rangle + b - y_{\mathrm{i}} & \leq \epsilon + {\xi^*}_{\mathrm{i}} \\ \xi_{\mathrm{i}}, {\xi^*}_{\mathrm{i}} \quad \geq 0 \end{cases} \tag{1.10}$$

In SVR, two types of "slack variables", ξ_{i} and ${\xi^*}_{\mathrm{i}}$ are used, for positive and negative deviations from the real value of the data points. The Lagrange multipliers, α_{i} are used to reformulate the optimization problem as:

$$f(x) = \sum_{i}\left(\alpha_{\mathrm{i}} - {\alpha^*}_{\mathrm{i}}\right)\langle x_{\mathrm{i}}, x \rangle + b \tag{1.11}$$

The deviations allowed by the two slack variables define the so called an ϵ-tube, within which the training data points have $(\alpha_{\mathrm{i}} - \alpha^{*}{}_{\mathrm{i}}) = 0$. The support vectors of SVR are associated with non-zero $(\alpha_{\mathrm{i}} - \alpha^{*}{}_{\mathrm{i}})$ and lie outside of this tube as shown in **Figure 1.7**.

1.5 Chemical Space Visualization

Chemical space represents the set of all chemically feasible compounds.[74] The systematic exploration of the chemical universe containing theoretically possible molecules (10^{60})[75] is an unfeasible task. However, relevant drug-like subspaces are rather distributed over small regions of the chemical space and computational visualization techniques are usually designed to focus on those regions. Unlike support vector machine models, which have a typical black-box behavior, visualization techniques are more intuitive and interpretable. Computational chemical space visualization techniques can be mainly classified into coordinate-based and coordinate-free representations.

1.5.1 Coordinate-based representation

In coordinate-based representation, numerical descriptors are used to generate a coordinate system onto which molecules are projected based on their descriptor values. The proximity of molecules in this type of representation corresponds to their similarity.

Activity landscape is one of the preferred coordinate-based models used for chemical space visualization, mainly focusing on SAR analysis. It is graphical representation that systematically integrate structural similarity and potency difference between compounds to highlight distinct SAR features among compounds sharing same specific biological activity.[58] Activity landscape used in SAR analysis is derived from chemical spaces, in which, distance relationships reflect structural similarity between compounds. Hence, compounds that are close to each other represent structural analogs. The position of the compounds are annotated with potency information.

The underlying chemical space used in activity landscape modeling is generally high-dimensional, and thus statistical dimensionality reduction techniques such as principal component analysis (PCA)[74,75] and multidimensional scaling (MDS)[76] must be applied to generate interpretable 2D or 3D activity landscape. PCA generates a set of non-correlated vectors called principal components by orthogonal transformation of the original multi-dimensional descriptor space. The resulting principal components are ranked in decreasing order of the variance observed in the original high dimensional space, i.e., the first principal component accounts for the highest variability in the data than the succeeding ones. MDS is a technique for non-linear mapping of data points from initial high dimensional space to a lower dimension in a way that preserves the distance between objects in the original space as much as possible. Although dimensionality reduction techniques are typically accompanied by loss of information they are indispensable for the challenging task of navigating through high dimensional chemical spaces.

3D activity landscape[77] resemble actual geographical landscape, thus providing an intuitive way to visualize activity landscapes. A high-dimensional fingerprint-based chemical space is first reduced to 2D space by applying PCA or MDS. An activity surface is then added as a third dimension to the 2D projection and colored based on the potency values of the compounds. Hence, significant potency differences between compounds can be clearly identified in 3D activity landscape. Depending on the underlying SAR characteristics of the compounds, activity landscape can be either smooth and easily traversed, or may have rugged surfaces. Continuous SARs correspond to smooth regions or gently rolling hills of activity landscape. In contrast, SAR discontinuity is represented by rough regions of activity landscape. Continuous and discontinuous SARs are not mutually exclusive, and hence, they may also coexist. The combination of both continuous and discontinuous phenotypes in a single data set is termed heterogeneous SARs. The landscape resulting from heterogeneous SARs is further termed variable activity landscape.[78]

Generative topographic mapping

Generative topographic mapping (GTM)[5] is a probabilistic dimensionality reduction approach that has been applied for visualization as well as development of predictive models. GTM was introduced by Bishop et al[5] as a method that allows non-linear mapping of objects from a multi-dimensional space to a 2D space. The probabilistic topology preserving nature of GTM has made it a popular tool in chemical space visualization and analysis. In addition, GTM-based predictive models have been successfully used for classification and regression purposes.

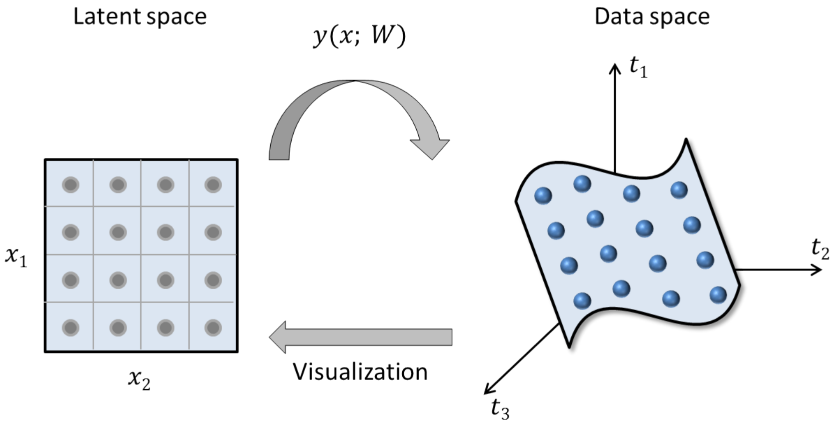


Figure 1.8: Generative topographic mapping. Non-linear mapping of data points from multi-dimensional data space to 2D latent space using GTM. A manifold is placed in original descriptor space to capture the data probability distribution via an ensemble of Gaussian functions. The Gaussian functions situated on the manifold are obtained by non-linear transformation $y(x; W)$ from the grid nodes in the latent space. The molecules, assigned with a specific probability to each node of the manifold, are finally projected onto the latent space for visualization.

The GTM algorithm places a two dimensional manifold akin of a "rubber sheet" into an initial high-dimensional space populated by "frame set" compounds. The manifold is twisted and turned in the original descriptor space to optimally accommodate a maximum of "frame set" compounds in order to reproduce the best data distribution via simulated probability distribution function. This function comprises an ensemble of Gaussian functions located at the nodes of a 2D grid related to the manifold. The extent to which the manifold may be distorted is controlled by a set of predefined parameters. The molecules are assigned with a specific probability to each node of the manifold, which is then projected onto a 2D latent space representing the final GTM landscape. Unlike regular 2D landscapes, in case of GTM, molecules are distributed such that each molecule is associated with one or more nodes. **Figure 1.8** presents a schematic representation of non-linear mapping using GTM.

A responsibility matrix $R(M, K)$ is calculated to record the probability of a given molecule M being assigned to node K. It is also represented as the responsibility of node K towards molecule M. Since a molecule is associated to one or more nodes of a GTM, its responsibility is also distributed over several nodes and the overall probability to find a molecule anywhere on the map is given by $\sum R(M, K)$, which sums up to 1.0. Furthermore, a set of compounds S residing in node K is defined by cumulative responsibilities of K towards all of its members, i.e. $\rho(S, K) = \sum R(M, K)$. The cumulative responsibility is the node bound density distribution of all members of a set of compounds and represents the fuzzy membership of these compounds in a given node of GTM.

In the following, generation and validation of GTM manifold using ISIDA-GTM tool[34,79,80] is described in brief. A set of different parameters govern the manifold construction, including the choice of initial descriptor space, "frame set" compounds and GTM setup parameters. As systematic exploration of all the parameter combinations is computationally expensive, an evolutionary algorithm[81] was used in our studies to search for optimal parameter setting. A chromosome vector encodes all required parameter settings as follows:

1. ISIDA descriptors defining initial high dimensional chemical space, each of which was considered as a possible choice.

2. Number of radial basis functions (RBF) defining the manifold (varied between 4 and 30).

3. RBF width (varied between 0.2 and 2.5).

4. Regularization coefficient (varied between 10^{-2} and 10^{2}).

5. Method of prediction (k-nearest neighbor or Bayesian modeling) to be used for classification models obtained from the manifold. The choice of prediction method therefore does not affect the resulting manifold.

6. Frame set compounds defining the reference frame for fitting the manifold in initial descriptor space.

The compounds in "frame set" are not labeled with any property annotation as they are used only to define regions of chemical space through which the manifold traverses. It is reasonable to have more than one frame set such that the GTM algorithm can span several relevant chemical space zones. The manifold construction is therefore unsupervised as no information about the property of the "frame set" compounds are taken into account.

Once the manifolds are obtained, they are validated using a different set of compounds termed "selection set". Instances in "selection set" are labeled with their respective class labels (for classification models) or property values (for regression models). GTM-based models are generated for the "selection set" compounds and conventional p-fold (p=3) cross validation is performed. The property value or class label of one-third of the compounds in this set are predicted using models trained on remaining two-third of the compounds. The prediction performance is evaluated by calculating different measures of accuracy for regression and classification tasks.

In case of regression, root mean square error and cross validated determination coefficient are calculated.[82]

$$RMSE = \sqrt{\frac{\sum\limits_{i=1}^{n}(pred_{\mathrm{i}} - exp_{\mathrm{i}})^2}{n}} \tag{1.12}$$

$$Q^2 = 1 - \frac{\sum\limits_{i=1}^{n}(exp_{\mathrm{i}} - pred_{\mathrm{i}})^2}{\sum\limits_{i=1}^{n}(exp_{\mathrm{i}} - \overline{exp})^2} \tag{1.13}$$

where, $pred_i$ and exp_i are the predicted and experimental values of the i^{th} compound, n is the total number of compounds, and $\overline{exp}$ is the mean over all experimental values.

Since, GTM-based classification[79,80] are only suitable for binary classification of given instances the balanced accuracy is determined as follows:

$$BA = \frac{0.5 \quad TP}{TP + FN} + \frac{0.5 \quad TN}{TN + FP} \tag{1.14}$$

where, *TP, TN, FP,* and *FN* are true positives, true negatives, false positives, and false negatives, respectively.

Average and standard deviation are calculated for root mean square deviation, determination coefficient, and balanced accuracy values originating from each cross validation experiment. The final fitness score of the map is calculated by taking this average score, penalized by the standard deviation.

Compound assignment to the nodes of GTM is defined by its responsibility vector. As GTM is complaint with neighborhood principle,[83] compounds with similar responsibility vectors lie close to one another in GTM landscape and have similar values of the underlying molecular descriptors. A good quality GTM landscape can hence be used to identify and extract regions of chemical space where similar compounds reside. These compounds should have similar values of responsibility vector, which suggests grouping of compounds based on their responsibility vectors. However, it is unlikely to find two or more compounds with identical responsibility vectors because they are real valued vectors. A cell-based partitioning of the responsibility vector space is applied to cluster compounds with similar responsibility vectors. Each responsibility vector is converted to a discretized form called responsibility pattern (RP)[84] as follows:

$$RP(M,K) = [10 \times R(M,K) + 0.9] \tag{1.15}$$

As such, if the responsibility of molecule M to node K is below 0.01, its contribution to the RP is zero. Values larger than 0.01 are binned at intervals of 0.1, increasing the contribution by +1 to the corresponding RP value. i.e. RP(M,K) = 1 if $0.01 \leq R(M,K) < 0.11$, RP(M,K) = 2 if $0.11 \leq R(M,K) < 0.21$ and so on. RPs thus provide an organizing principle for compounds on GTM as each RP defines cluster of related compounds which can be further analyzed.

1.5.2 Coordinate-free representation

Coordinate-based representation suffers from a number of caveats, including the "curse of dimensionality".[85] Application of dimensionality reduction techniques and consequent loss of information is unavoidable to achieve interpretable low dimensional representation of original high-dimensional chemical space. In contrast, coordinate-free representation, as the name suggests do not require construction of any coordinate system and dimensionality reduction methods. Inter-compound distances in coordinate-free representation do not have any chemical meaning.

Network-like similarity graph (NSG)[86] is a coordinate-free visualization technique that provides access to global and local SAR features in data sets. NSG is similarity-based network representation, in which nodes correspond to individual compounds that are connected by edges if predefined similarity criteria is met. Similarity is measured by calculating Tanimoto[54] coefficient from molecular fingerprints. Nodes are colored with respect to the potency of the corresponding compounds. Further, node size indicates a score accounting for local SAR discontinuity contribution of the compound.[86,87] Larger nodes therefore represent compounds whose potency value significantly differ from their immediate structural neighbors. An extension of NSG was introduced to graphically visualize and identify "selectivity cliffs", defined as pairs of structurally similar compounds with distinct selectivity profiles, i.e., selective for two different targets.[88] The nodes of selectivity landscape are annotated by ratio of compound potency values against two targets under investigation. Further extensions of NSG include multi-target activity landscape[89] and molecular mechanism-based NSG.[90]

Bipartate matching molecular series (BMMS) graph[91] is an SAR network based on MMS relationships of a given data set. The network representation comprises two types of nodes, key nodes and molecule nodes. Key nodes represent the MMP core that is shared by a given MMS whereas individual compounds are represented by molecule nodes and are color-coded according to their potency values. Molecule nodes representing molecules of individual MMS are connected to the corresponding key nodes. BMMS graph allows systematic organization of data set compounds based on their substructural relationships and highlights disjoint subgraphs that correspond to specific structural modifications revealing important SAR patterns.

In contrast to global SAR visualization, similarity potency tree (SPT)[92] is compound-centric representation, and hence, reveals local SAR features with respect to a selected reference compound. SPT is a tree-like graph that incorporates structural similarity and potency relationships among a set of structural analogs. The root node of the tree structure corresponds to the reference compound and is connected by edges to its nearest structural neighbors. The structural similarity to the reference compound decreases towards the leaf nodes. Horizontal and vertical graph reading of SPT facilitates meaningful interpretation of SAR information in compound subsets.[92]

Coordinate-free representations capturing hierarchical substructure relationships are based on molecular scaffolds. Intuitive network for structure-activity relationship analysis (inSARa)[93] is bipartite network representation like BMMS graph that capture hierarchical maximum common substructure (MCS) relationships. Layered skeleton-scaffold organization (LASSO) graph[94] organizes compounds in form of compound-scaffold-carbon skeleton hierarchy. BM scaffold and CSK are generated for individual compounds. Compounds sharing the same BM scaffold are organized into a pie chart, such that the slices of the pie are colored according to the their potency. Further, topologically equivalent BMS (sharing a common CSK) are grouped together into boxes that are connected if they share a substructure and layered in hierarchical fashion.[94] AnalogExplorer, designed for graphical exploration of large analog series, enables systematic organization and comparison of structurally related or distinct analog series in heterogeneous data sets.[95]

Chemical Space Network

Chemical space network[85] (CSN) was introduced by Maggiora and Bajorath in 2014 as a generalized form of similarity-based compound networks, with a principal focus on visualization of biologically relevant chemical spaces.

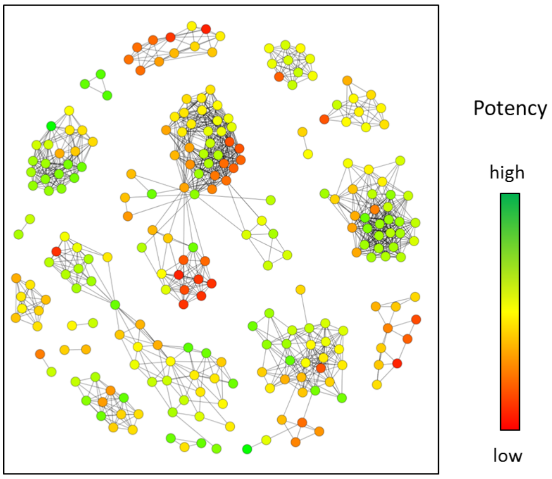


Figure 1.9: Chemical space network. Chemical space network generated is shown. Nodes correspond to compounds and are colored by their potency value going from red (weakly potent) through yellow to green (highly potent). Compounds meeting predefined structural similarity criteria are connected by edges.

In CSN, nodes represent compounds and edges indicate the similarity relationship between two nodes. If a predefined similarity relationship between two compounds is fulfilled, the corresponding nodes are connected by an edge. Nodes are annotated with compound potency. The CSNs that have been reported so far differ mainly with respect to the similarity criteria employed during their construction. While numerical similarity measures such as Tanimoto coefficient were used to generate threshold CSN (THR-CSN),[96,97] others utilized substructure-based similarity assessment like MMP (MMP-CSN)[97] and MCS (MCS-CSN).[98] In THR-CSN, similarity between compounds are continuous numerical values and edges are drawn if a predefined threshold criteria is met. In contrast, similarity assessment in substructure-based CSN rather takes a binary form (similar vs. not similar) and is independent of threshold values. Therefore, substructure-based CSN is more intuitive from medicinal chemistry perspective. The CSNs discussed so far represent undirected networks, since the similarity measures are generally symmetric in nature.[85] A recently introduced Tv$_{\text{MCS}}$-CSN[99] utilizes a hybrid similarity measure based on Tversky coefficient[54] and MCS to generate an asymmetric variant of CSN. In another interesting extension of CSN, distance relations in high-dimensional property spaces are transformed into similarity relations to generate so-called trans-CSN.[100]

Structure-Activity Relationship Matrix

SAR matrix[101] (SARM) resembles R-group table and organizes structural analogs in matrix-like form based on MMP and MMS relationships. Is is designed for efficient extraction of SAR patterns from groups of structurally related bioactive compounds of large data sets in chemically interpretable manner. SARM is constructed following a two-step MMP generation procedure based on Hussain and Rea[30] fragmentation scheme.

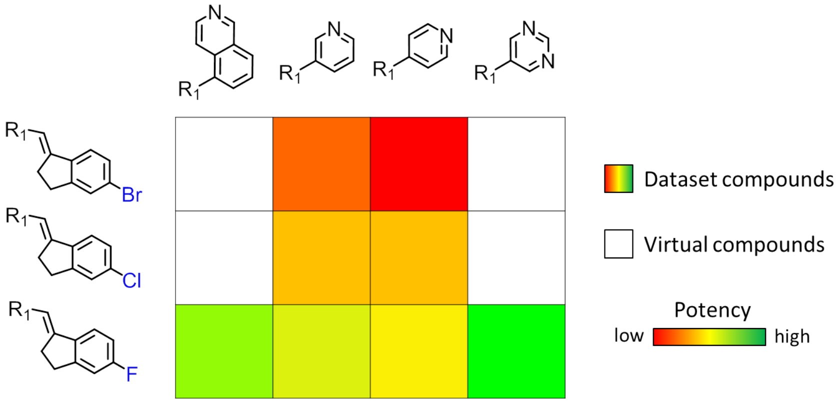


Figure 1.10: SAR matrix SAR matrix data structure is shown. Rows and columns represent MMS core and substituents, respectively. Colored cells represent real compounds present in the data set and colored based on their potency value following the color scheme used in Figure 1.9. Non-colored cells correspond to virtual compounds.

The first stage MMPs are generated by fragmenting all compounds and the resulting key (MMP core) and value fragments are organized in an index table to identify MMS. In the second stage, each MMP core obtained from the first round of fragmentation is again subjected to MMP generation, giving rise to a second set of key and value fragments (key' and value'). MMSs are then grouped together as “matching MMSs”, if they share a common key’. The key’ and value’ of “matching MMSs” are finally arranged along the rows and columns of SARM, respectively. Each row of a SARM represents an MMS with a unique core, each column represents a substituent and each cell represents a compound. Cells can either correspond to real compounds from the data set and are colored according to their potency or virtual compounds that do not exist. The potency-based coloring of cells can also be replaced by any other desired property of the compound.

1.6 Thesis Outline

This thesis comprises seven representative studies that are organized into individual chapters as follows. *Chapter 2* explores structural diversity of scaffolds obtained from publicly available bioactive compounds. Structurally distinct scaffold combinations were systematically extracted and their potency range distributions were monitored globally as well as at individual target level. In *Chapter 3*, we demonstrate the application of GTM to detect privileged structural motifs for three major pharmaceutical target families. Responsibility pattern-based compound subsets extracted from GTM landscape were prioritized in individual target families and subjected to RECAP-fragmentation and subsequent visual analysis to identify privileged substructures. A two-layered SAR visualization tool, combining GTM and CSN methodology is introduced in *Chapter 4*, with an application for detection of interesting SAR patterns in very large and heterogenous data sets. GTM provided access to the “global view” of the activity landscape from which regions of high SAR information were extracted. SAR trends were explored in CSNs of SAR-informative local molecular communities obtained from GTM to analyze structural modifications resulting in potency changes. In *Chapter 5*, SVM-based classification and regression models were used for predicting activity cliffs. In addition to predicting activity cliff or non-cliff status of MMPs, the directionality of potency changes were predicted. A large-scale analysis of ionization state distribution of bioactive compounds is reported in *Chapter 6*. Ionization state preference across target sets and target superfamilies were studied. Relationships between ionization state class and potency of compounds were analyzed. *Chapter 7* introduces star coordinates, a tool for visualization of compounds in a multi-property landscape. Numerically equivalent solutions arising from multi-objective optimization tasks were compared in star coordinates to identify drug-like subspaces in multi-dimensional property space. A diagnostic tool to monitor SAR progression in lead optimization is introduced in *Chapter 8*. SARMs were systematically generated for evolving sets of compounds from public repositories and lead optimization settings. Further, SARMs were classified and annotated with two numerical scores to evaluate their SAR information at different time points. Finally, *Chapter 9* summarizes the major outcomes of the thesis with concluding remarks.

Chapter 2

Structural Diversity and Potency Range Distribution of Scaffolds from Compounds Active Against Current Pharmaceutical Targets

Introduction

The scaffold concept is popular in pharmaceutical research for describing the core structure of active compounds. Two major areas in which scaffolds have been extensively applied are privileged substructure identification and scaffold hopping. Scaffolds are often associated with biological activity of compounds to determine core structures that are essential to specifically bind to a given pharmaceutical target. Such core structures can provide starting point for compound optimization. Another interesting application of scaffolds is in identification of structurally distinct active molecules with similar potency. Drug discovery projects usually involve optimization of multiple compound series. This is done mainly to ensure the availability of structurally distinct candidate compounds sharing similar activity, in case one or more compounds need to be discarded due to liability issues. Scaffold hopping techniques are usually applied in such settings to extrapolate from a given structural series to another with similar biological activity. The success of computational scaffold hopping methods in part relies on the structural diversity of available scaffolds.

Given the increase in the number of compounds with reported structure-activity data, the systematic organization and analysis of their scaffolds has become more challenging. In addition, researchers have defined scaffolds in many different ways and a concrete definition of scaffolds hardly exists. In our analysis, we have used a widely applied scaffold definition by Bemis and Murcko, so-called BM scaffolds. Assessing structural similarity between BM scaffolds is non-trivial given the varying degree of structural relationships they may share. A pair of BM scaffolds that differs only by a single hetero-atom is structurally more similar compared to the one that differs by the number of rings. To account for differing levels of similarity among BM scaffolds, a further abstraction of the structure was obtained by generating cyclic skeletons (CSKs). A BM scaffold represents a set of structurally related compounds and a CSK represents two or more topologically equivalent BMs. The hierarchical organization of molecular structures from compounds to BM to CSKs, facilitated comparison of structures at different levels of abstraction.

A comprehensive scaffold survey has been reported in this work, with an aim to systematically determine the structural diversity of compounds globally as well as at individual target level. Scaffolds and CSKs were systematically extracted from compounds with high-confidence activity data in ChEMBL. Combinations of structurally distinct CSKs and scaffolds were identified following a set of well-defined structural rules. Scaffolds and CSKs were assigned to potency intervals based on the potency range distribution of compounds represented by them.

Research Article

Future Medicinal Chemistry

For reprint orders, please contact reprints@future-science.com

Structural diversity and potency range distribution of scaffolds from compounds active against current pharmaceutical targets

Aim: Large numbers of biologically active compounds are available from which scaffolds (core structures) can be isolated and compared focusing on structural, potency and promiscuity criteria. **Results:** A computational analysis has been carried out to characterize all scaffolds and cyclic skeletons contained in currently available compounds from medicinal chemistry sources. Compounds active against hundreds of pharmaceutical targets were found to contain many structurally distinct scaffolds and cyclic skeletons. For given targets, these scaffolds often represent highly potent compounds. **Conclusion:** There is an abundance of scaffold diversity among specifically active compounds indicating that many pharmaceutically relevant proteins are highly permissive small molecular targets. These findings have several implications for drug discovery and design.

Shilva Kayastha[1], Dilyana Dimova[1], Dagmar Stumpfe[1] & Jürgen Bajorath*,[1]

[1]Department of Life Science Informatics, B-IT, LIMES Program Unit Chemical Biology & Medicinal Chemistry, Rheinische Friedrich-Wilhelms-Universität, Dahlmannstr. 2, D-53113 Bonn, Germany

*Author for correspondence:
Tel.: +49 228 2699 306
Fax: +49 228 2699 341
bajorath@bit.uni-bonn.de

Scaffold analysis

The **scaffold** concept is applied in medicinal chemistry to describe core structures of active compounds [1]. Scaffolds can be defined in different ways and the term is often rather loosely used in literature [1]. According to a widely applied definition following a molecular hierarchy [2], scaffolds are defined as follows: *"The scaffold is obtained from a compound by removal of all (nonring) R-groups while retaining all ring structures and linker fragments between ring structures."*

From these so-called Bemis-Murcko (BM) scaffolds (also termed frameworks), following the original publication [2], one can further abstract from chemical structure by generating so-called **cyclic skeletons** (CSKs) [3], defined as follows: *"The CSK is obtained from a scaffold by converting all heteroatoms to carbon and setting all bond orders to one."*

BM scaffolds might display very different structural relationships, in other words, they might be structurally distinct or very similar (e.g., only distinguished by a single heteroatom in a ring or a bond order). This wide spectrum of scaffold structures and their varying relationships often complicates the assessment of scaffold diversity [1]. As a further abstraction from scaffolds, CSKs are, by definition, topologically distinct (although they also span a wide range of structural relationships). According to the underlying molecular hierarchy, each unique BM scaffold represents a set of compounds (sharing the same scaffold) and each CSK a set of topologically equivalent scaffolds. A BM scaffold or CSK is often designated a **chemotype**, another term used in medicinal chemistry to describe a compound class with a specific core structure (which is also rather loosely applied).

The popularity of the scaffold concept in medicinal chemistry can at least in part be attributed to the search for **privileged substructures** [4] or target class-directed **masterkeys** [5], in other words, core structures that preferentially interact with a given target family such as serine proteases or kinases. Such privileged structural motifs are typically explored at the level of molecular scaffolds. Another popular application of the scaffold concept is the **scaffold hopping** exercise [6] that has become a hallmark of virtual screening efforts [7,8]. Scaffold hopping refers to the ability of computational methods to recognize specifically

FUTURE SCIENCE part of fsg

10.4155/FMC.14.156 © 2015 Future Science Ltd ISSN 1756-8919

Key terms

Scaffold: Representation of the core structure of a compound. Scaffolds can be generated in different ways, for example, by removal of R-groups from compounds or by following retrosynthetic rules. The term is often synonymously used with framework.

Cyclic skeleton: A further abstraction from scaffolds. Cyclic skeletons are derived from scaffolds by converting all heteroatoms to carbon and setting all bond orders to one.

Chemotype: Defined here as a chemical core structure representing a class of compounds.

Privileged substructures/masterkeys: Rationalized in medicinal chemistry as core structures that are recurrent in compounds active against a given target family. Privileged structural motifs are thought to exhibit selectivity for a target family, but not specificity for individual family members.

active compounds with different scaffolds; in other words, to extrapolate from a given structural series or class and identify another sharing similar activity. A variety of computational and chemical scaffold hopping approaches have been introduced [8,9]. These include, for example, computational pharmacophore and shape searching or fingerprint methods to identify compounds with core structures that are increasingly dissimilar to those of given reference molecules. If no computational methods are applied in medicinal chemistry, scaffold hopping is mostly attempted on the basis of prior knowledge or chemical intuition. The assessment of the scaffold hopping potential of computational methods is also affected by the wide spectrum of structural relationships among scaffolds [9]. For example, many formal scaffold hops might involve very similar structures (and are hence much easier to facilitate than others).

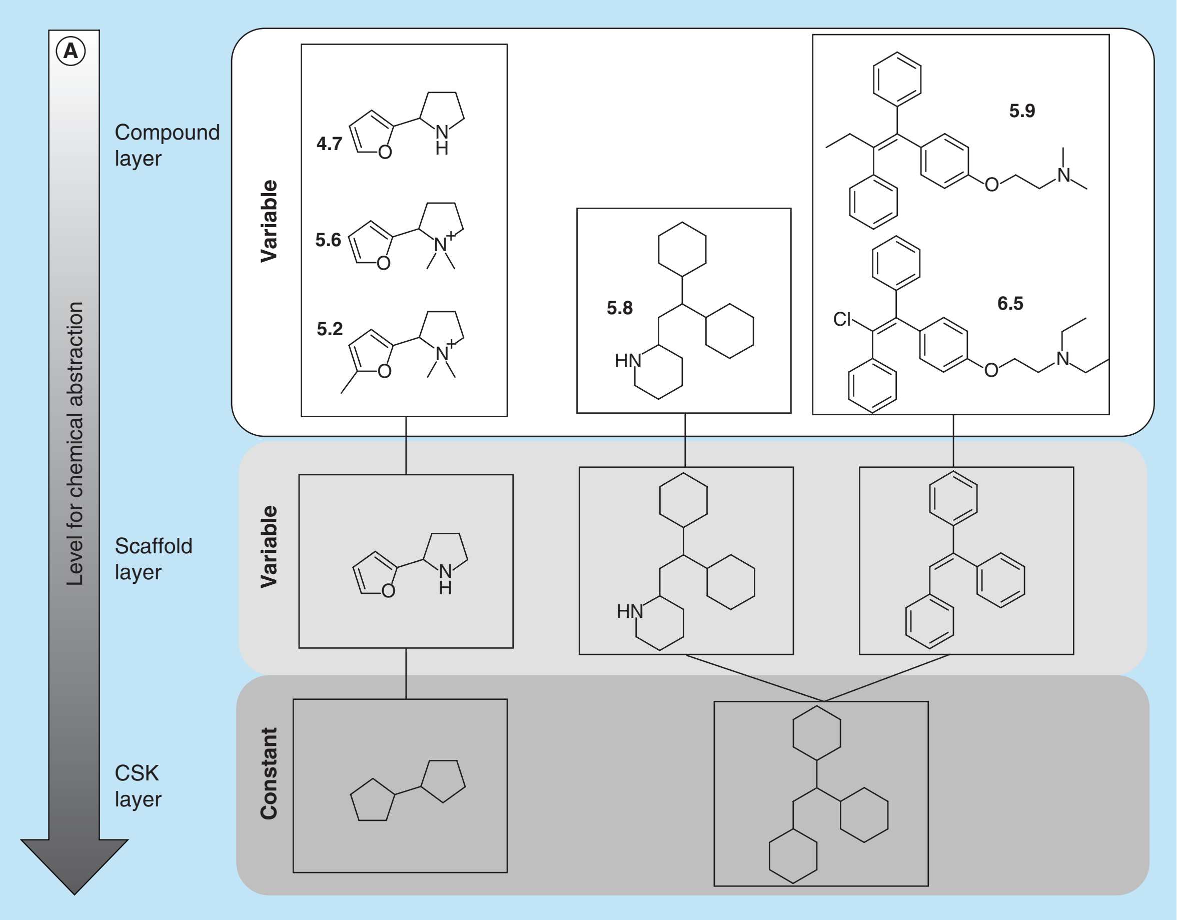


Figure 1. Levels of chemical abstraction and analysis workflow. (A) Levels of chemical abstraction. The hierarchical relationship between CSKs, scaffolds and compounds is illustrated. The CSK layer considered in our analysis consists of CSK combinations, the scaffold layer of scaffold combinations and the compound layer of compound pairs. A combination of SD CSKs often covers multiple scaffolds. Similarly, a combination of structurally distinct scaffolds can represent multiple compounds. In this example, the CSK combination covers three distinct scaffolds and six compounds that form two scaffold combinations and nine compound pairs.
comb.: Combinations; cpds: Compounds; CSK: Cyclic skeleton; poss.: Possible; pot: Potency; SD: Structurally distinct; SS: Structurally similar.

Different methodologies have been introduced for the organization and graphical analysis of scaffold populations following molecular hierarchies including scaffold trees and hierarchical graph representations [10–12]. Among these are approaches to explore structure–activity relationships (SARs) at the level of scaffolds and predict new active compounds [11–13]. Furthermore, scaffold surveys have been carried

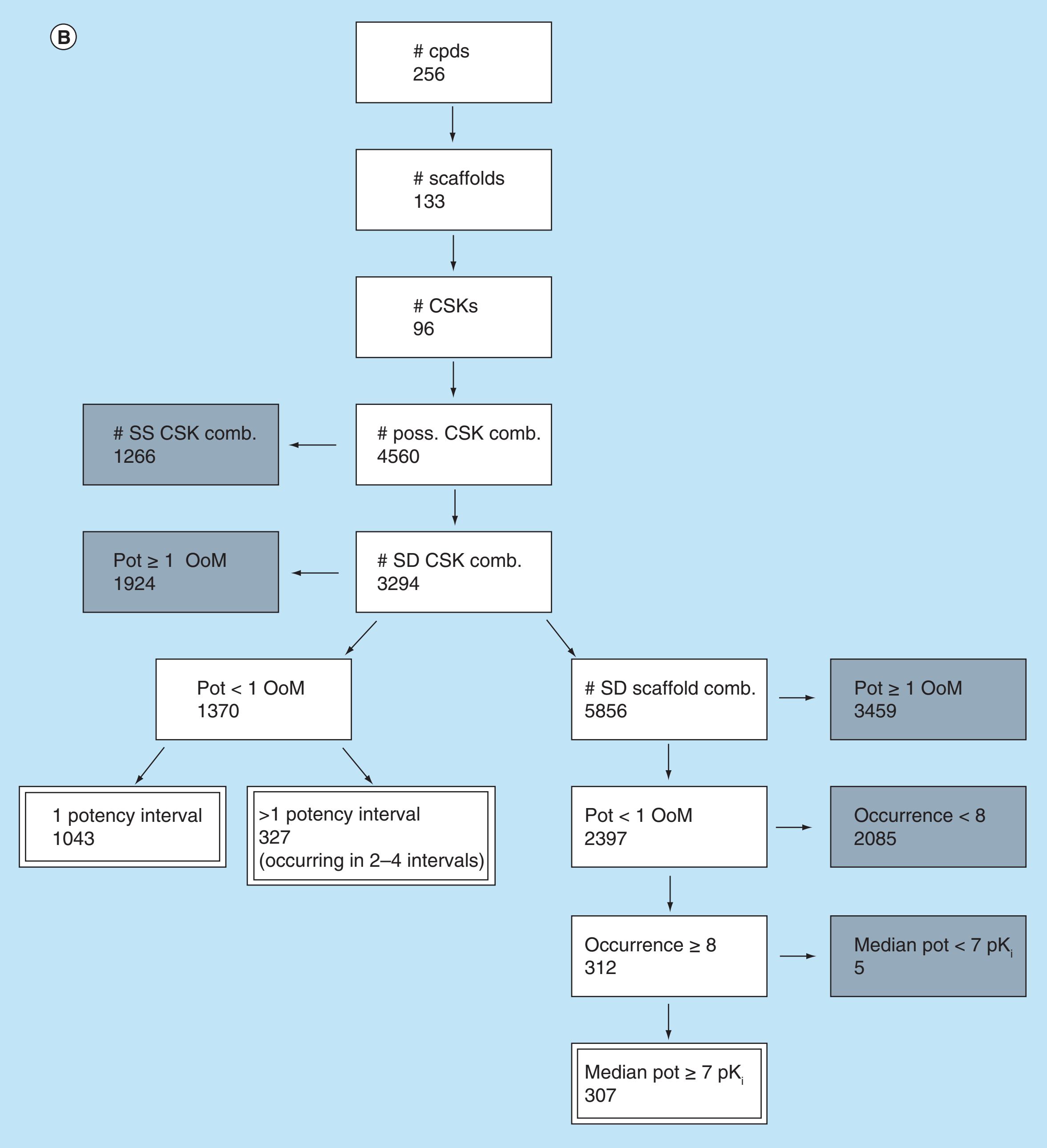


Figure 1. Levels of chemical abstraction and analysis workflow (cont.). **(B)** Analysis workflow for an exemplary target set. The analysis workflow is reported using neurokinin 1 receptor ligands as an example. The number of compounds, scaffolds and CSKs is reported. Furthermore, the number of CSK and scaffold combinations that satisfy (white) or do not satisfy (gray) different filtering criteria is given. Structurally similar CSK combinations and CSK combinations representing pairs of compounds with at least tenfold potency difference were excluded. In addition, scaffold combinations representing less than eight compound pairs were not further considered. Finally, combinations formed by scaffolds representing compounds in low potency intervals were excluded from the analysis (see the methodological sections for further details). Numbers of all CSK and scaffold combinations that were considered for further analysis are reported in double bordered boxes.
comb.: Combinations; cpds: Compounds; CSK: Cyclic skeleton; poss.: Possible; pot: Potency; SD: Structurally distinct; SS: Structurally similar.

Key terms

Scaffold hopping: Term introduced in the context of computational (virtual) compound screening to describe the ability of computational methods to identify specifically active compounds containing different scaffolds.

Promiscuity: Defined as the ability of small molecules to specifically interact with multiple biological targets.

out for organic molecules [14], bioactive compounds [15] and drugs [16–18] and to compare different chemical libraries [19,20]. In one of these studies, scaffolds were isolated from sets of compounds that were active against different targets [15]. In this case, for 70 targets, 50 or more scaffolds were found in known ligands, hence providing a first view of scaffold diversity among specifically active compounds.

In this study, we have systematically extracted scaffolds and CSKs from compounds with high-confidence activity data available in ChEMBL [21,22], the major public repository of compounds from medicinal chemistry sources. Well-defined structural relationships between CSKs and scaffolds were explored and combinations of structurally distinct CSKs and scaffolds were identified. The potency range distribution of compounds represented by scaffolds and corresponding CSKs hierarchy was determined and scaffolds/CSKs were assigned to corresponding potency intervals. Our analysis has revealed that structurally distinct scaffolds representing highly potent compounds frequently occur for unexpectedly large numbers of pharmaceutical targets, which has several implications for drug discovery and design.

Compound data mining

Our analysis was facilitated with in-house written Perl and Python scripts for compound retrieval and scaffold/CSK generation and KNIME workflows [23,24] for scaffold/CSK analysis. From ChEMBL (version 17), compounds with activity annotations for human targets were assembled and organized into individual targets sets (with each set consisting of all compounds active against a given target). We have only considered compounds with precisely defined assay-independent equilibrium constants (K_i values) reported at the highest confidence level (ChEMBL confidence score 9). Compounds with multiple activity annotations for the same target were only selected for further analysis if all values fell within one order of magnitude (then, the average potency value was calculated as the final activity annotation). Depending on its degree of **promiscuity**, a compound might be assigned to multiple target sets.

From all qualifying compounds, BM scaffolds were isolated and for all scaffolds, the corresponding CSKs were generated. Only target sets containing at least ten different BM scaffolds were retained for further analysis. On the basis of these selection criteria, a total of 72,254 compounds with activity against 317 different targets were obtained (forming 317 target sets). These compounds yielded a total of 31,638 scaffolds and 20,536 CSKs over all target sets (counting multiple occurrences) and comprised 16,899 and 8704 unique scaffolds and CSKs, respectively. Out of all 31,638 scaffold instances, 22,067 (69.7%) represented only a single active compound while the remaining 9571 scaffolds represented multiple compounds (on average, 5.2 compounds per scaffold). Furthermore, of all 20,536 CSKs, 15,300 (74.5%) represented a single scaffold and the remaining 5236 CSKs multiple scaffolds (on average, 3.1 scaffolds per CSK).

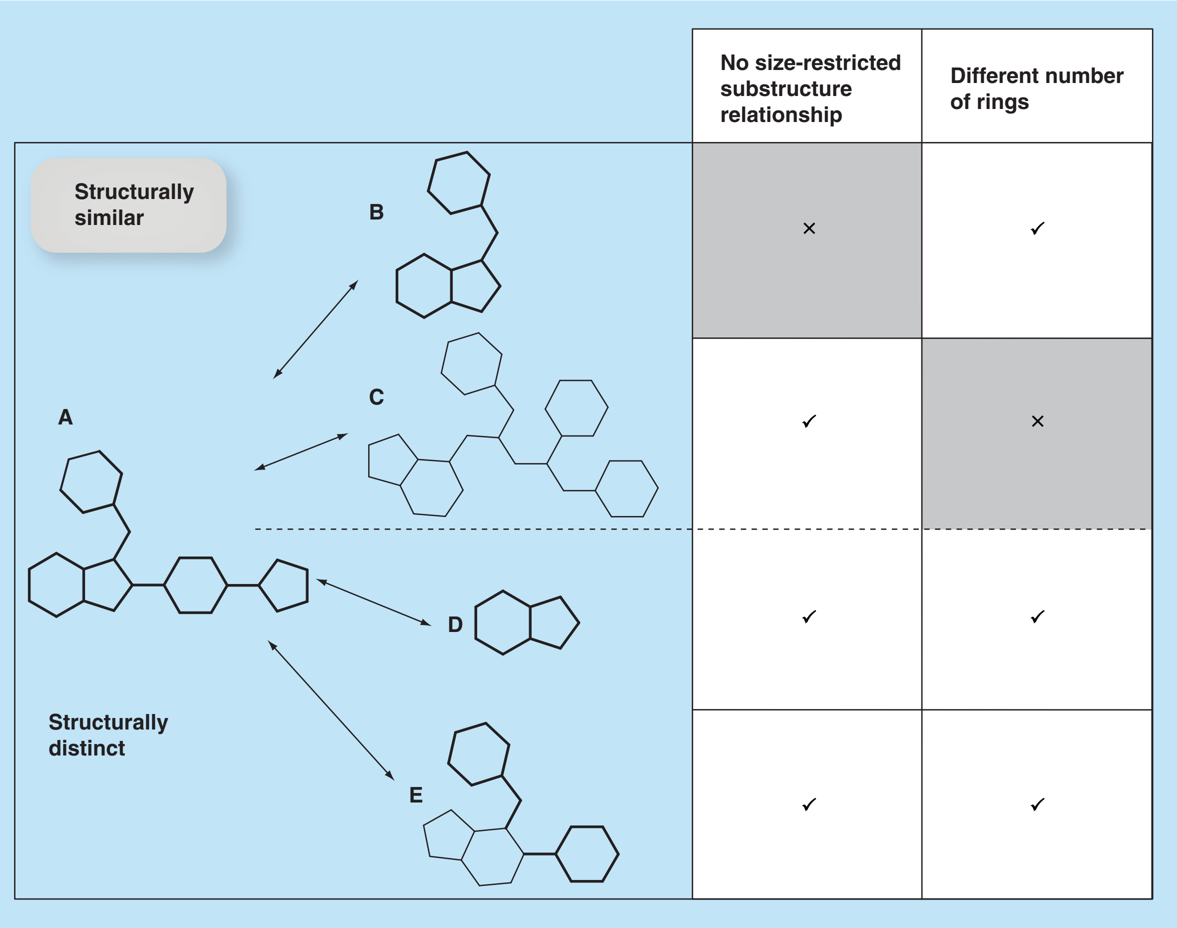
Scaffold & cyclic skeletons combinations

For all scaffolds and CSKs, pairwise comparisons were systematically carried out and all possible combinations of different scaffolds (scaffold layer) and CSKs (CSK layer) were obtained, as illustrated in Figure 1A. Each CSK combination covered varying numbers of scaffold combinations. Each scaffold in a given combination was required to yield a different CSK. The resulting scaffold and CSK layer contained a total of 4,837,969 and 2,325,241 unique scaffold and CSK combinations, respectively. Figure 1B summarizes the analysis protocol and provides a representative example.

Structural relationships

We focused our analysis on structurally distinct scaffolds and CSKs. Hence, all enumerated CSK combinations were evaluated according to predefined structural rules, as illustrated in Figure 2. Differences in topology and ring numbers, substructure relationships and CSK size restrictions were considered as criteria to identify combinations of structurally distinct CSKs. Combinations of different CSKs were categorized as ‘structurally distinct’ or ‘structurally similar’. The latter category was not further explored in our analysis.

Two CSKs were considered to be structurally distinct if they contained different number of rings and if they were not involved in a substructure relationship (i.e., a CSK was not a substructure of another). CSKs with substructure relationships were only considered structurally distinct if the size (number of carbon atoms) of one CSK was at least twice the size of the other CSK, as illustrated in Figure 2. In addition, CSKs/scaffolds consisting of only a 6-membered ring (e.g., benzene) were removed because of their very large frequency of occurrence. By definition, combinations of structurally distinct CSKs meeting these criteria exclusively covered combinations of structurally distinct scaffolds. Because all structural comparisons were carried out in a pairwise manner, CSK and scaffold pairs represented



No size-restricted substructure relationship	Different number of rings
×	✓
✓	×
✓	✓
✓	✓

Figure 2. Identification of structurally distinct cyclic skeleton combinations. Shown are two structurally similar (A↔B, A↔C) and two structurally distinct cyclic skeleton combinations (A↔D, A↔E). The underlying structural rules are summarized.

the basic data unit for our analysis. Pairs of structurally distinct CSKs or scaffolds according to the criteria detailed above are designated in the following as 'structurally distinct pairs'.

A total of 1,737,599 structurally distinct CSK and 4,340,548 scaffold combinations were obtained from compounds active against 315 targets.

Potency range assignment

For each CSK and scaffold, potency values from all corresponding compounds were assembled and the CSKs and scaffolds were assigned to potency intervals covering the entire potency range populated with preselected ChEMBL compounds. Each CSK and scaffold was assigned to potency intervals into which at least one of its compounds fell. Hence, a CSK and scaffold might be assigned to different intervals, and counted several times, depending on the potency distribution among its compounds. Often, multiple compounds represented by a CSK or scaffold were observed to fall into the same potency range. Out of all 31,638 scaffolds, approximately 80% represented compounds falling into a single potency interval. The remaining 6465 scaffolds represented compounds with potency values falling into multiple intervals, with the number of intervals ranging from two to ten. In addition, of all 20,536 CSKs, approximately 69% covered compounds in a single potency interval, and the remaining 6433 CSKs represented compounds with potency values in multiple intervals (with the number of intervals again ranging from two to ten). This global distribution resulted from summarizing potency range assignments that were separately carried out for each individual target set. In Table 1, the potency range distribution of compounds, scaffolds and CSKs over all target sets is reported. As can be seen, large numbers of CSKs and scaffolds were assigned to the submicro- to subnanomolar potency subrange, in particular, into the intervals ranging from [7–8) to [9–10), because of the represented highly potent compounds.

Structurally distinct CSK & scaffold combinations with similar potency

Next, CSK and scaffold combinations were assigned to potency intervals. For each CSK or scaffold

Table 1. Potency range distribution.

Potency range	Number of CPDs	Number of scaffolds	Number of CSKs	Number of TIDs
[12–13)	5	4	4	4
[11–12)	62	49	47	24
[10–11)	551	368	301	81
[9–10)	3819	2292	1777	196
[8–9)	13,142	7094	5142	264
[7–9)	18,696	10,089	7352	293
[6–7)	18,249	10,341	7792	310
[5–6)	13,053	7629	5972	301
[4–5)	3531	2241	1831	229
[3–4)	907	584	489	131
[2–3)	153	101	88	32
[1–2)	69	44	32	9
[0–1)	17	14	11	5

The table reports the global potency range distribution of compounds, scaffolds, CSKs and TIDs. A potency interval [X-Z) includes all pK_i values Y, with X ≤ Y < Z. Compounds, CSKs and scaffolds were counted multiple times if they occurred in multiple target sets and/or multiple potency intervals. Scaffolds and CSKs were assigned to a given interval if there was at least one corresponding compound with a pK_i value falling into the interval. A target set was assigned to a given interval if it contained at least one compound with a pK_i value falling into the interval.
CPD: Compound; CSK: Cyclic skeleton; TID: Target set.

combination, corresponding compounds were required to have potency values within one order of magnitude (less than tenfold potency difference). The mean potency of all qualifying compounds represented by a pair of CSKs or scaffolds was calculated to assign the pair to a potency interval. A total of 1,079,281 structurally distinct CSK and 2,422,712 scaffold combinations with similar potency were identified. These combinations accounted for 62.1 and 55.8% of all structurally distinct CSK and scaffold combinations, respectively.

The number of compound pairs represented by each structurally distinct scaffold combination was also determined. Scaffold combinations representing at least eight compound pairs were frequently observed.

Table 2. Distribution of structurally distinct cyclic skeleton/scaffold combinations.

Potency range	Number of CSKs	Number of SD_CSK_Cs	Number of scaffolds	Number of SD_Scaffold_Cs	Number of TIDs
[12–13)	4	2	4	2	2
[11–12)	44	55	51	67	8
[10–11)	540	2111	669	3386	51
[9–10)	3286	45,637	4484	84,621	158
[8–9)	7763	216,479	11,247	443,507	236
[7–8)	1,1240	428,702	16,534	890,000	266
[6–7)	12,066	478,902	17,122	973,618	290
[5–6)	9180	266,303	12,306	486,939	272
[4–5)	3788	28,596	4845	45,075	192
[3–4)	637	1741	777	2478	85
[2–3)	83	172	95	229	11
[1–2)	29	67	34	91	4
[0–1)	4	2	6	4	2

The table reports the potency range distribution of SD_CSK_Cs and SD_Scaffold_Cs and the CSKs and scaffolds involved in these combinations. For each potency interval, the number of TIDs with at least one SD_CSK_C (and hence at least one SD_Scaffold_C) falling into the interval is given.
CSK: Cyclic skeleton; SD_CSK_C: Structurally distinct CSK combination; SD_Scaffold_C: Structurally distinct scaffold combination; TID: Target set.

In total, 259,101 of such combinations were identified in 238 target sets.

Table 2 reports the potency range distribution of all combinations over all target sets (the analysis was again separately carried out for each individual target set). Essentially, the entire global potency range was populated with varying numbers of structurally distinct CSK and scaffold combinations. The majority of combinations fell into the intervals ranging from [5–6) to [8–9) (submicromolar range).

High-potency CSK & scaffold combinations

Table 2 also reveals the presence of many combinations in the submicro- to subnanomolar potency intervals from [7–8) to [9–10). A total of 690,818 structurally distinct CSK and 1,418,128 scaffold combinations were identified in these intervals as high-potency combinations, which represented compounds active against 269 targets. From these combinations, a total of 13,138 CSKs and 19,932 scaffolds were obtained. Thus, there were surprisingly large numbers of scaffolds available for many targets that represented highly potent compounds. In 40.3% of all scaffold combinations, the median potency for both scaffolds was at least submicromolar (i.e., pK_i of 7 or greater; pK_i represents the logarithmic value of the equilibrium constant). These scaffold combinations were detected in 259 target sets.

Figure 3 shows the potency range distribution of structurally distinct CSK combinations for exemplary target sets (i.e., antagonists of the related neurokinin 1 and 2 receptors). These sets were of moderate size consisting of 256 and 212 compounds, respectively. Differences in the potency range distribution become immediately apparent. For neurokinin receptor 2, most

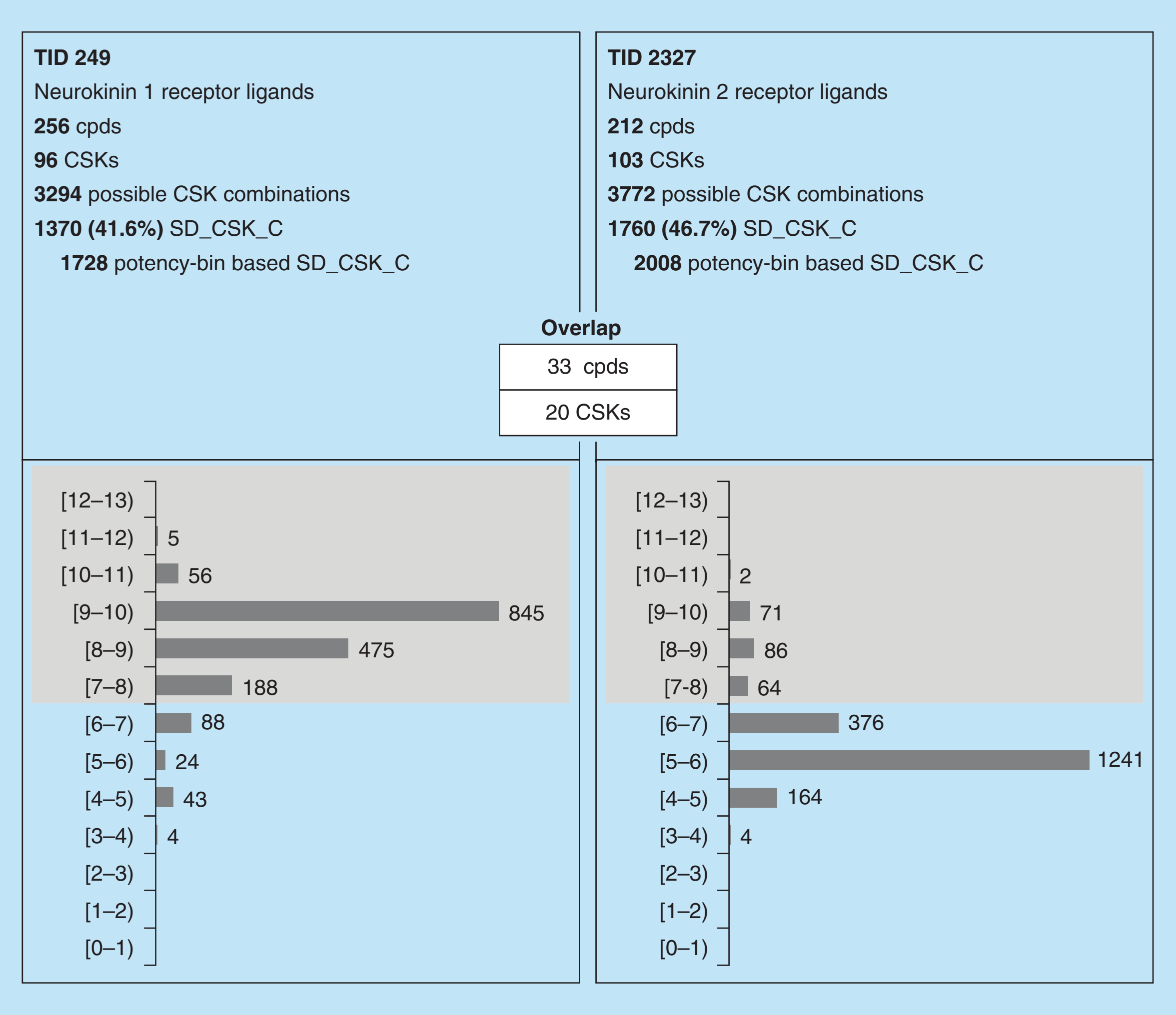


Figure 3. Targets with comparable distribution of cyclic skeleton combinations. For two exemplary target sets (ligands of the neurokinin 1 and 2 receptor), the number of compounds, CSKs as well as SD_CSK_Cs is given. In addition, the number of overlapping compounds and CSKs is given. The distribution of SD_CSK_Cs over the global potency range (pK_i intervals) is reported in bar plots.

cpds: Compounds; CSK: Cyclic skeleton; SD_CSK_C: Structurally distinct CSK combination; TID: Target set.

Table 3. Top ten target sets with largest numbers of high-potency cyclic skeletons.

Target name	Number of SD_CSK_Cs	Number of CSKs	Number of SD_CSK_Cs [HP]	Number of CSKs [HP]
Histamine H3 receptor	89,453	512	75,285	475
μ-opioid receptor	81,862	537	48,582	410
Κ-opioid receptor	73,473	516	44,639	392
Dopamine D2 receptor	98,936	556	40,241	380
Serotonin 1a (5-HT1a) receptor	52,035	381	38,068	341
Cannabinoid CB2 receptor	75,477	462	40,806	334
δ-opioid receptor	63,472	468	33,559	326
Dopamine D3 receptor	52,374	419	31,373	321
Adenosine A2a receptor	101,474	543	27,539	306
Adenosine A3 receptor	80,068	482	29,238	297

Reported are the top ten target sets (represented by target names) with largest numbers of CSKs in high-potency intervals ([HP]) involved in the formation of SD_CSK_Cs, in other words, from [7–8) to [12–13). For each target set, the total number of CSKs and SD_CSK_Cs across the entire potency range and the number of CSKs [HP] and SD_CSK_Cs [HP] are reported.
CSK: Cyclic skeleton ; HP: High potency; SD_CSK_C: Structurally distinct CSK combination.

of the CSK combinations were assigned to low-potency intervals. By contrast, for neurokinin receptor 1 most combinations fell into high-potency intervals. For example, 845 combinations represented compounds in the nano/subnanomolar interval [9–10). These combinations involved 58 CSKs that represented 78 scaffolds and 154 compounds.

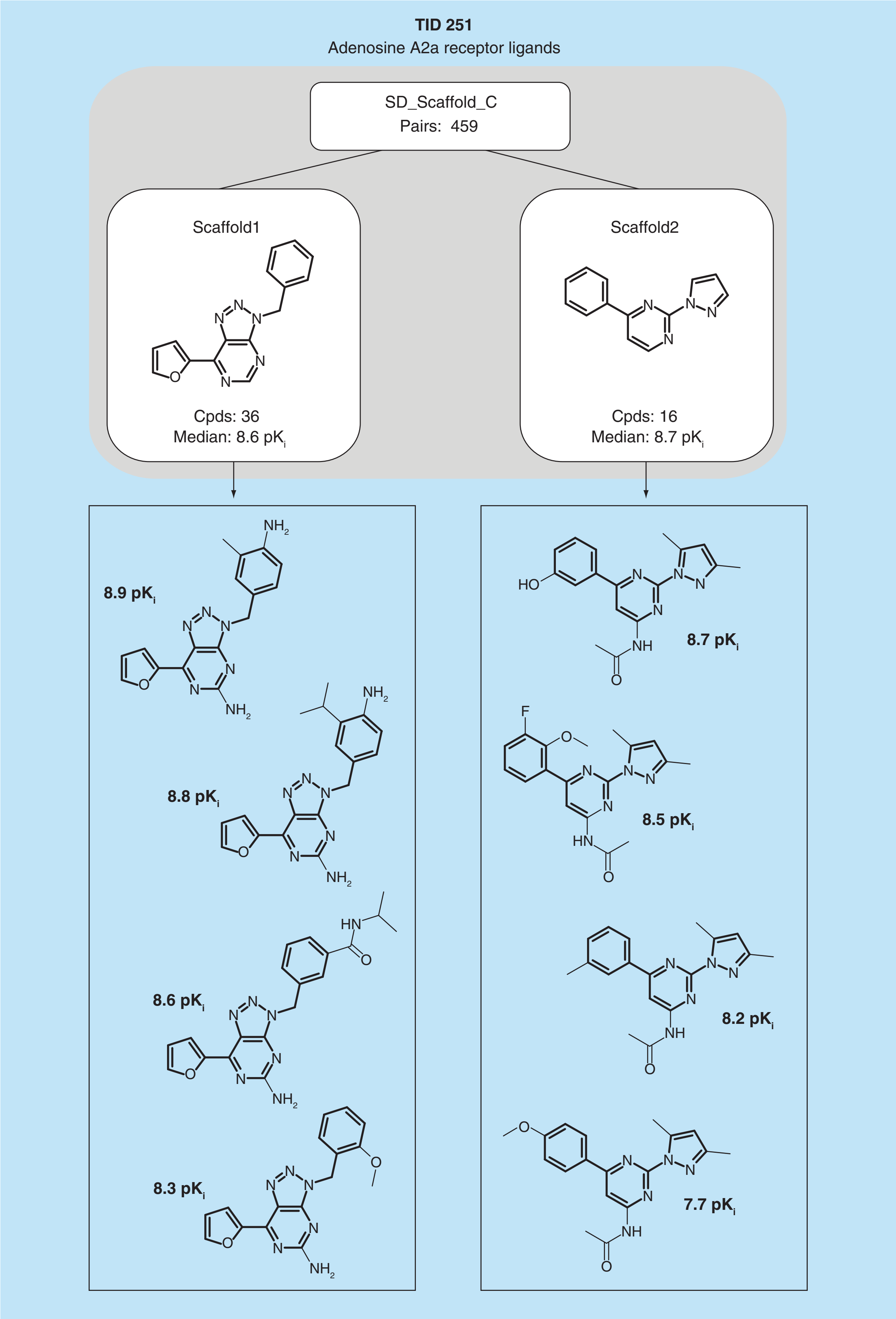
Table 3 reports the target sets containing the largest numbers of high-potency CSKs. All of these sets consisted of compounds active against various G-protein-coupled receptors. Other target sets with large numbers of high-potency CSKs included, for example, serine proteases and potassium channels. Figure 4 shows an example of a high-potency scaffold combination for ligands (antagonists) of the adenosine A2a receptor. These two structurally distinct scaffolds represented 36 and 16 compounds, respectively, yielding median potency values in the low nanomolar range.

Furthermore, 117,436 scaffold combinations were identified in 187 target sets that were found at least eight times in compounds with high potency (i.e., pK_i of 7 or greater). These combinations contained a total of 13,016 scaffolds (representing multiple counts for different potency intervals). Most of the scaffold combinations (100,835) occurred in a single target set (i.e., the compounds represented by a given scaffold were only active against a single target). Only 7628 combinations were detected in multiple (2–5) target sets. Figure 5 shows examples of dual-target scaffold combinations. For example, the scaffold combination in Figure 5A represented 42 compounds that were antagonists of the nociceptin and μ-opioid receptors. These compounds formed 301 pairs shared by these two target sets. The compound pairs displayed a different potency distribution for the two targets because they had overall higher (low nano- to subnanomolar) potency against the nociceptin receptor. Comparable observations were made for the scaffold combination in Figure 5B that represented 91 pairs of 32 antagonists shared by the serotonin 2A and 2B receptors.

Conclusion

We have carried out a comprehensive analysis of CSKs and scaffolds from the currently available spectrum of bioactive compounds, following the compound-scaffold-CSK hierarchy. Scaffolds and CSKs were calculated to represent core structures of compounds. The major goal of our study has been to systematically determine how structurally diverse compounds with specific target activities are and how difficult it might be to identify different compound classes with a desired target activity. Our analysis is not the first scaffold survey reported, but probably the most comprehensive one to date. Moreover, at least two criteria set it apart from previous studies. First, we have focused the analysis on CSKs and scaffolds that are structurally distinct on

Figure 4. Structurally distinct scaffold combination (see facing page). An example of a structurally distinct scaffold combination (SD_Scaffold_C) from an adenosine A2a receptor ligand set is shown. The number of qualifying compound pairs (occurrences) yielding this combination is reported. In addition, for each scaffold, the number of compounds involved in this SD_Scaffold_C and their median potency is given. Representative compounds are shown. The scaffolds are highlighted.
cpds: Compounds; SD_Scaffold_C: Structurally distinct scaffold combination; TID: Target set.


TID 251
Adenosine A2a receptor ligands
SD_Scaffold_C
Pairs: 459
Scaffold1
Cpds: 36
Median: 8.6 pK_i
Scaffold2
Cpds: 16
Median: 8.7 pK_i
8.9 pK_i
8.8 pK_i
8.6 pK_i
8.3 pK_i
8.7 pK_i
8.5 pK_i
8.2 pK_i
7.7 pK_i

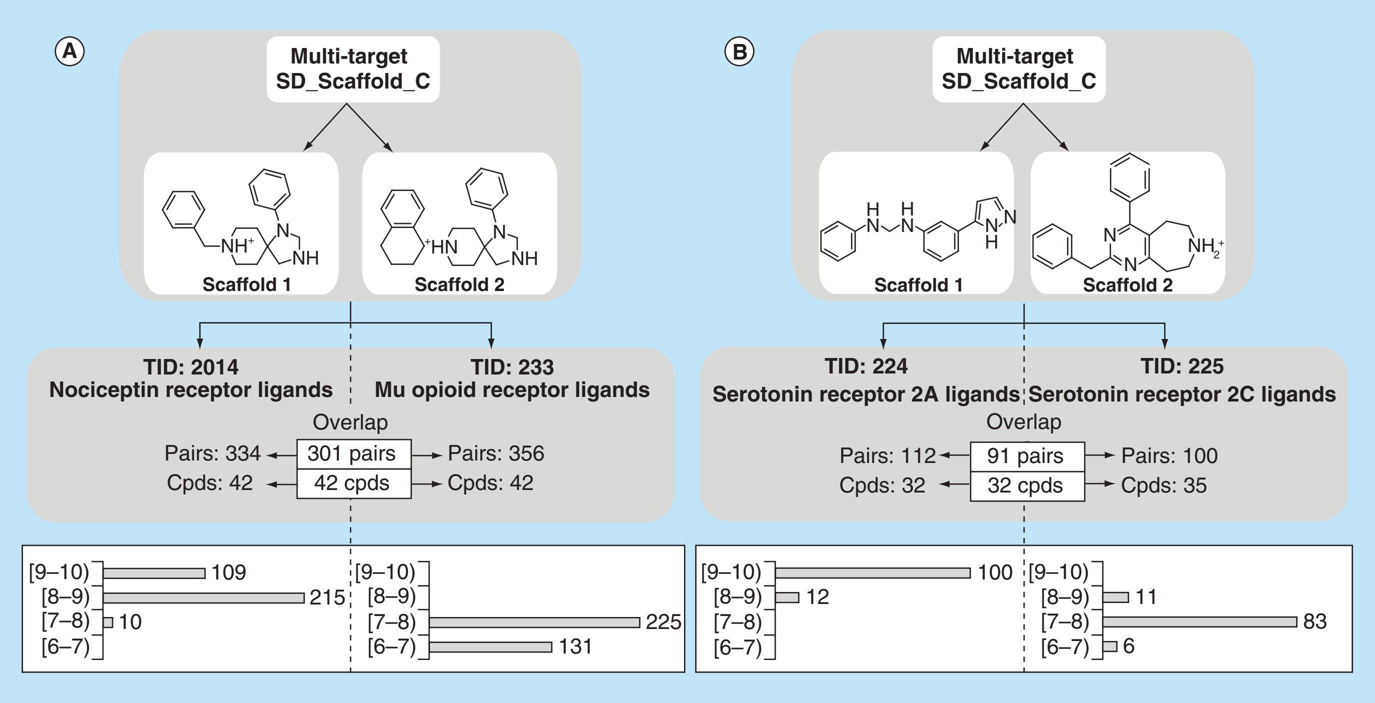


Figure 5. Structurally distinct dual-target scaffold combinations. Two structurally distinct scaffold combinations (SD_Scaffold_Cs) are shown that represent compounds with dual-target activity. For each target set, the number of qualifying compound pairs (occurrences) and compounds involved in these pairs is reported. The overlap between these sets (compounds and pairs with dual-target activity) is given in white boxes. In addition, the target-dependent potency range (pK_i) distribution of all compound pairs is reported (bottom). Dual-target SD_Scaffold_Cs originate from the **(A)** nociceptin/μ-opioid receptor and **(B)** serotonin 2A/2C receptor ligand sets.
SD_Scaffold_C: Structurally distinct scaffold combination; TID: Target set.

the basis of topological, substructure relationship and molecular size criteria. Second, we have assigned CSKs and scaffolds to potency intervals on the basis of the compounds they represent. In addition, we have exclusively used high-confidence activity data and (assay-independent) equilibrium constants as potency measurements to ensure a high level of data integrity. From more than 70,000 qualifying compounds with activity against more than 300 targets, a pool of approximately 17,000 unique scaffolds and approximately 9000 CSKs was obtained. The scaffold-to-CSK ratio was low, only approximately 2:1, indicating the presence of many different scaffolds. However, it was unknown how many of these scaffolds (and CSKs) might structurally be similar or analogous. By focusing the analysis on pairs of structurally distinct CSKs and scaffolds, we have identified 1,079,281 CSK and 2,422,712 scaffold pairs representing compounds active against 315 targets. These structurally distinct pairs yielded a total of 20,272 and 31,116 CSKs and scaffolds, respectively (multiple counts for different target sets). Thus, there were unexpectedly large numbers of scaffolds and CSKs involved in the formation of structurally distinct pairs for a large number of targets. These findings indicated that many pharmaceutical targets were capable of interacting with highly diverse compounds. For scaffolds and CSKs from structurally distinct pairs, the scaffold-to-CSK ratio was approximately 1.5:1, thus even smaller than the ratio calculated for target sets (see above). Moreover, when analyzing the potency range distribution of CSKs and scaffolds, we found that many structurally distinct scaffold pairs represented highly potent compounds. Although one might perhaps expect that structurally diverse compounds active against a given target might often be weakly potent (assuming that only a limited number of compounds could be optimized to yield high potency), the results of our analysis provide a different picture. In fact, we have found that multiple high-potency CSKs and scaffolds originating from structurally distinct pairs were available for 269 targets with, on average, 74 scaffolds per target. Each of these scaffolds represented, on average, 2.2 compounds, typically with potency in the nanomolar range. Thus, for the majority of targets, high-potency scaffolds were detected that were structurally distinct from others.

Taken together, the results of our analysis make it possible to draw several conclusions. Many pharmaceutically relevant proteins are excellent small molecular targets for which structurally distinct and highly potent compounds are already available. Hence, it is likely that structurally distinct compound series can be developed as alternative drug candidates for such targets. Furthermore, scaffold hopping exercises are likely to yield novel compounds in many instances. If a variety of structurally distinct scaffolds representing active compounds already exists, it should not be too difficult to identify additional ones. Moreover, the availability of limited compound information for targets might not necessarily mean that they are poor small molecular targets (although this might also be true depending on the particular target). For many popular drug targets, biologically relevant chemical space is most likely not yet comprehensively explored. Hence, more chemical diversity would need to be considered in trying to identify novel chemical entities for targets for which only one or a few active chemotypes are currently available. There should be more structurally diverse active compounds to identify.

To aid in further analysis all scaffold sets and related information are freely available from the authors upon request.

Future perspective

Analyzing compound activity data following the compound-scaffold-CSK hierarchy makes it possible to explore core structures of bioactive compounds and their structural relationships. Furthermore, activity data can be taken into consideration. An increasing level of structural abstraction is required to assign compounds to structural classes and organize them with respect to biological activities. Therefore, a systematic scaffold and CSK organization also provides a basis for the annotation of core structures with other biological properties, in addition to potency and the generation of bio-chemical profiles of scaffolds with defined structural relationships. For example, metabolic stability, availability, pharmacokinetic and/or toxicology data might be collected for compounds represented by structurally distinct scaffolds sharing the same specific activity. The generation of such high-level profiles might aid in the selection of preferred compound classes for given targets. This should provide interesting opportunities for future research, especially in the pharmaceutical industry where a wealth of biological and pharmacological data is typically available (as opposed to academia). Annotated compound-scaffold-CSK hierarchies should provide valuable resources for compound selection and design, especially in light of our findings that many pharmaceutical targets are capable of recognizing structurally distinct compounds. Hence, a search for alternative compound classes with desirable property profiles might often be a meaningful approach in the course of discovery projects. In addition, another opportunity for future research is provided by attempting to systematically complement a structural organization of scaffolds with chemical reaction information. This is an important aspect for compound development because molecular hierarchies generated in the context of scaffold analysis usually do not take reaction information into account. It follows that scaffolds often have different synthetic accessibility. Thus, complementing molecular hierarchies with reaction information should be an important step forward for the practice of medicinal chemistry.

It is conceivable that the scaffold-CSK universe representing currently explored biologically relevant chemical space could be integrated with high-level bio-chemical profiles, reaction schemes and target as well as structure–activity relationship information to yield advanced ontologies for future pharmaceutical research; a fairly ambitious, but not unrealistic goal.

Acknowledgement

We thank the OpenEye [25] Free Academic Licensing Program for providing a free academic license for the chemistry

Executive summary

Compound data mining

- A comprehensive analysis of cyclic skeletons (CSKs) and scaffolds in current bioactive compounds has been carried out.

Scaffold & CSK combinations

- Structural rules have been formulated and applied to focus on structurally distinct pairs of CSKs and scaffolds.
- For hundreds of pharmaceutical targets, large numbers of structurally distinct CSK and scaffold combinations have been identified.

Potency range assignment

- In many instances, structurally distinct CSKs and scaffolds represented highly potent compounds.
- On the basis of these findings, many targets are permissive to structurally diverse highly potent compounds.

High-potency scaffold & CSK combinations

- A limited number of promiscuous scaffolds have been identified that yield highly potent compounds with activity against multiple targets.

Conclusion

- For many targets, it should be possible to evolve structurally distinct compound series into highly potent candidates.
- Computational scaffold hopping exercises might often be less challenging than generally assumed.

toolkit that was used to aid in the generation of data mining routines.

Financial & competing interests disclosure

D Stumpfe is supported by Sonderforschungsbereich 704 of the Deutsche Forschungsgemeinschaft. The authors have no other relevant affiliations or financial involvement with any organization or entity with a financial interest in or financial conflict with the subject matter or materials discussed in the manuscript apart from those disclosed.

No writing assistance was utilized in the production of this manuscript.

References

Papers of special note have been highlighted as:
• of interest

1 Hu Y, Stumpfe D, Bajorath J. Lessons learned from molecular scaffold analysis. *J. Chem. Inf. Model.* 51(8), 1742–1753 (2011).

2 Bemis GW, Murcko MA. The properties of known drugs. 1. Molecular frameworks. *J. Med. Chem.* 39(15), 2887–2893 (1996).

- **Introduction of a widely used scaffold definition.**

3 Xu Y-J, Johnson M. Algorithm for naming molecular equivalence classes represented by labeled pseudographs. *J. Chem. Inf. Model.* 41(1), 181–185 (2001).

4 Evans BE, Rittle KE, Bock MG *et al.* Methods for drug discovery: development of potent, selective, orally effective cholecystokinin antagonists. *J. Med. Chem.* 31(12), 2235–2246 (1988).

- **Origin of the privileged substructure concept.**

5 Müller G. Medicinal chemistry of target family-directed masterkeys. *Drug Discov. Today* 8(15), 681–691 (2003).

6 Schneider G, Neidhart W, Giller T *et al.* “Scaffold-hopping” by topological pharmacophore search: a contribution to virtual screening. *Angew. Chem. Intl Ed.* 38(19), 2894–2896 (1999).

- **Introduction of the scaffold-hopping idea.**

7 Walters WP, Stahl MT, Murcko MA. Virtual screening - an overview. *Drug Discov. Today* 3(4), 160–178 (1998).

8 Geppert H, Vogt M, Bajorath J. Current trends in ligand-based virtual screening: molecular representations, data mining methods, new application areas, and performance evaluation. *J. Chem. Inf. Model.* 50(2), 205–216 (2010).

9 Hongmao S, Tawa G, Wallqvist A. Classification of scaffold hopping approaches. *Drug Discov. Today* 17(7–8), 310–324 (2012).

10 Wilkens SJ, Janes J, Su AI. HierS: hierarchical scaffold clustering using topological chemical graphs. *J. Med. Chem.* 48(9), 182–193 (2005).

11 Schuffenhauer A, Ertl P, Roggo S *et al.* The scaffold tree--visualization of the scaffold universe by hierarchical scaffold classification. *J. Chem. Inf. Model.* 47(1), 47–58 (2007).

- **One of the original, and to this date, widely applied structural organization and visualization scheme for large scaffold sets.**

12 Gupta-Ostermann D, Hu Y, Bajorath J. Introducing the LASSO graph for compound data set representation and structure–activity relationship analysis. *J. Med. Chem.* 55(11), 5546–5553 (2012).

13 Wetzel S, Klein K, Renner S *et al.* Interactive exploration of chemical space with scaffold hunter. *Nat. Chem. Biol.* 5(8), 581–583, (2009).

14 Lipkus AH, Yuan Q, Lucas KA *et al.* Structural diversity of organic chemistry. A scaffold analysis of the CAS registry. *J. Org. Chem.* 73(12), 4443–4451 (2008).

15 Hu Y, Bajorath J. Global assessment of scaffold hopping potential for current pharmaceutical targets. *Med. Chem. Comm.* 1(5), 339–344 (2010).

16 Wang J, Hou T. Drug and drug candidate building block analysis. *J. Chem. Inf. Model.* 50(1), 55–67 (2008).

17 Hu Y, Bajorath J. Many drugs contain unique scaffolds with varying structural relationships to scaffolds of currently available bioactive compounds. *Eur. J. Med. Chem.* 76(1), 427–434 (2014).

18 Taylor RD, MacCoss M, Lawson ADG. Rings in drugs. *J. Med. Chem.* 57(14), 5845–5859 (2014).

- **Recent extensive survey of ring structures in drugs.**

19 Shelat AA, Guy RK. Scaffold composition and biological relevance of screening libraries. *Nat. Chem. Biol.* 3(8), 442–446 (2007).

20 Langdon SR, Brown N, Blagg J. Scaffold diversity of exemplified medicinal chemistry space. *J. Chem. Inf. Model.* 51(9), 2174–2185 (2011).

21 Gaulton A, Bellis LJ, Bento AP *et al.* ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res.* 40(database issue) D1100–D1107 (2012).

22 ChEMBL.
www.ebi.ac.uk/chembl.

23 Tiwari A, Sekhar AK.. Workflow based framework for life science informatics. *Comput. Biol. Chem.* 31(5), 305–319 (2007).

24 KNIME.
www.knime.org.

25 Openeye.
www.eyesopen.com.

Summary

A computational framework to characterize all scaffolds and cyclic skeletons from currently available bioactive compounds and analysis of their structural relationships and potency distribution has been presented. A total of 72,254 qualifying compounds with activity against 317 targets served as an initial pool from which 31,638 scaffolds and 20,536 CSKs were extracted. 69.7% of scaffolds represented a single compound and 74.5% of CSKs represented a single scaffold, providing a strong evidence that most of the scaffold and CSKs are not extensively explored. Each scaffold and CSK was assigned to potency intervals into which at least one of its compound fell. Most of them fell in high potency ranges, indicating the availability of many scaffolds derived from highly potent compounds.

All possible combinations of different scaffold and CSKs were generated and structural rules were defined to identify "structurally distinct pairs". Structural rules were based on comparisons of topology, ring numbers and substructure relationships with size restrictions. A total of 4,340,548 and 1,737,599 structurally distinct scaffold and CSK combinations were obtained, respectively. Individual "structurally distinct pair" was also assigned to potency intervals, if corresponding compounds had less than tenfold difference in their potency. A qualifying pair thus represents an instance of structurally distinct series of compounds sharing similar activity, a typical scaffold hopping instance. Of all the structurally distinct scaffold and CSK combinations reported above, 55.8% and 62.1%, respectively had similar potency, and thus, supposed to be of high scaffold hopping potential. Furthermore, 1,418,128 structurally distinct scaffold and 690,818 CSK combinations fell in high potency intervals of submicro- to subnanomolar ranges, i.e., from 7 $\mathrm{pK_i}$ to 9 $\mathrm{pK_i}$.

My major contributions to this work have been defining the structural rules for filtering structurally distinct scaffold combinations and identifying compounds represented by them, globally as well as in individual targets. Also, I have carried out the frequency and potency distribution analysis of structurally distinct scaffold combinations.

Our analysis clearly indicates the existence of a large number of structurally diverse scaffolds representing highly potent compounds active against many pharmaceutical targets. The development of structurally distinct compound series to search for alternative drug candidates for such targets thus most likely will yield successful outcome. In the next chapter, we will focus on another important application of scaffold concept, identification of privileged substructures.

Chapter 3

Privileged Structural Motif Detection and Analysis Using Generative Topographic Maps

Introduction

Privileged substructures represent the core structural unit found frequently among compounds active against a given target family and hence are believed to play a major role in their biological activity. They are thought to be selective towards a given target family but show promiscuity within members of the target family. Identification of privileged substructures in a group of compounds provides useful insight into understanding the underlying molecular interactions responsible for their biological activity. Because of the target family-specific nature of privileged substructures, they can serve as promising starting points in design of novel active compounds for a given target. The search for privileged substructures has therefore gained a lot of interest over the years in compound library design and drug discovery.

Several methods have been applied for identification of privileged substructures and a number of privileged scaffolds have so far been reported for drugs as well as natural products. Most of the earlier studies were scaffold-centric, and focused mainly on exploring molecular core structures or scaffolds. However, with continuous increase in the number of newly synthesized compounds, single molecular framework might not be sufficient to define the entire target-family-privileged structure. Further, drugs or bioactive compounds have several functional as well as pharmacophoric groups that contribute to the total free energy of drug-receptor complex, consequently affecting their activity. Scaffolds that were designated as “privileged” were often also found to be active against targets from other families as more compounds containing these scaffolds were synthesized and tested. It can be reasoned that instead of a single scaffold, a group of structurally related or specifically substituted similar scaffolds might be associated with molecules preferentially interacting with a given target family.

In this study, we demonstrated the application of GTM to prioritize compounds representative of target superfamilies and subsequently extract privileged structural motifs from them. Compounds projected onto GTM landscape were grouped together in the same or nearby nodes if they were structurally related, i.e., if they have similar responsibility vectors. Regions of the map that were preferentially populated by compounds active against specific target superfamily were extracted to explore any common substructures they shared. One of the major goals of the current study is to assess GTM from medicinal chemistry viewpoint in its ability to highlight emerging SAR patterns. A significant difference of this approach from most of the traditional scaffold-based techniques is that motifs here were extracted directly from compound clusters obtained from GTM without any prior knowledge of their classification.

Journal of Chemical Information and Modeling

Article

pubs.acs.org/jcim

Privileged Structural Motif Detection and Analysis Using Generative Topographic Maps

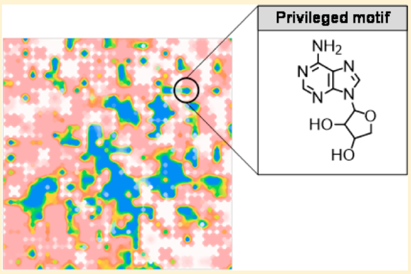
Shilva Kayastha,[†,‡] Dragos Horvath,[‡] Erik Gilberg,[†,§] Michael Gütschow,[§] Jürgen Bajorath,[*,†] and Alexandre Varnek[*,‡]

[†]Department of Life Science Informatics, B-IT, LIMES Program Unit Chemical Biology and Medicinal Chemistry, Rheinische Friedrich-Wilhelms-Universität, Dahlmannstr. 2, D-53113 Bonn, Germany

[‡]Laboratoire de Chemoinformatique, UMR 7140, Université de Strasbourg, 1 rue Blaise Pascal, Strasbourg 67000, France

[§]Pharmaceutical Institute, University of Bonn, An der Immenburg 4, 53121 Bonn, Germany

Supporting Information



ABSTRACT: Identification of "privileged structural motifs" associated with specific target families is of particular importance for designing novel bioactive compounds. Here, we demonstrate that they can be extracted from a data distribution represented on a two-dimensional map obtained by Generative Topographic Mapping (GTM). In GTM, structurally related molecules are grouped together on the map. Zones of the map preferentially populated by target-specific compounds were delineated, which helped to capture common substructures on the basis of which these compounds were grouped together by GTM. Such privileged structural motifs were identified across three major target superfamilies including proteases, kinases, and G protein coupled receptors. Traditionally, the search for privileged structural motifs focused on scaffolds, whereas motifs were detected here without prior knowledge of compound classification in GTMs. This alternative way of navigating medicinal chemistry space further extends the classical, scaffold-centric approach. Importantly, detected motifs might also comprise fuzzy sets of similar scaffolds, pharmacophore-like patterns, or, by contrast, well-defined scaffolds with specific substituent patterns.

1. INTRODUCTION

The concept of "privileged substructures" was originally introduced by Evans,[1] referring to core structures that are recurrent in compounds active against a given target family and therefore associated with that biological activity. Privileged substructures are thought to be selective toward a given target family but not individual family members. Compound library design focusing on such target family-directed structural motifs would be promising for identifying novel active compounds. The search for privileged structures has therefore remained popular over the years in library design and drug discovery. Most of the earlier studies directly focused on exploring molecular core structures or scaffolds, and a number of privileged scaffolds have been proposed for drugs and natural products. However, privileged scaffolds were often also found to be active against targets from other families as increasing numbers of compounds containing the scaffold were synthesized and tested. From a mechanistic perspective, it is plausible to attribute a privileged status not to a single scaffold but rather to a group of structurally related scaffolds, hence representing a structural motif by a large ensemble that also covers some variations.[2] In addition to classical scaffold-based analysis, approaches using decision trees[3] have also been employed to identify substructures that distinguish actives from inactives within a given compound collection.[4,5] In the current study, we report the application of GTM[6] to prioritize compounds from a large pool of target family selective sets and subsequently identify privileged substructural motifs on the basis of these compounds.

GTM is a machine learning method that allows nonlinear mapping of a set of molecules from a multidimensional chemical space to two-dimensional space for its visualization and analysis. The grid defines a GTM-based landscape and comprises a number of nodes onto which compounds are projected. In addition to chemical space visualization, GTM has been successfully used to build and optimize predictive models for classification and regression analysis. The GTM-based models have revealed performance levels comparable to other state-of-the-art machine learning methods such as support vector machines,[7] naïve Bayes[8] classifiers, or random forests.[9] As an extension of the GTM methodology, Stargate GTM (S-GTM),[10] has been introduced to bridge between descriptor and activity spaces and enable mapping of compounds from one space into the other.

In our analysis, "universal" GTMs[11] obtained from an evolutionary map-growing procedure were used, as further discussed below. These maps represented a subset of GTMs built using an initial set of ISIDA molecular descriptors.[12−14] So far, reported proof-of-concept studies have focused on quantitative validation of active vs inactive discriminatory power of GTM

Received: March 2, 2017
Published: April 14, 2017

DOI: 10.1021/acs.jcim.7b00128

models for diverse biological targets not related to any of those used for map fitting. This work continues the external validation effort by assessing, for targets from different superfamilies, the ability of the maps to discriminate between ligand classes. Another major goal of the present work is analyzing GTM from a medicinal chemistry perspective and assessing emerging structure—activity relationships.

The GTM-based approach enabled prioritization of regions in chemical space that were most representative of a given target family and provided the opportunity to directly map compounds and compare their assignment to individual nodes. Mapping aided in the prioritization of compounds primarily directed against a given target family and provided a basis for the subsequent extraction and visualization of key structural patterns akin to privileged substructures.

2. MATERIALS AND METHODS

2.1. Compound Data Sets. Our web server for virtual screening[15] was updated with the ChEMBL[16] release 20 entries, serving as a large compound background set for analyzing ligands of different target superfamilies highlighted in various activity class landscapes (*vide infra*). Specifically, subsets of compounds with well-defined activity (K_i and/or IC_{50}) values against human targets were assembled from ChEMBL. Target and family information was curated by combining the classification schemes of ChEMBL and UniProt.[17] Compounds with no information concerning the year of entry into the database were discarded. This resulted in a total of 152,274 compounds active against 1464 targets.

2.2. Generative Topographic Mapping. GTM is a method of nonlinear mapping of data points from a multidimensional chemical space to a two-dimensional space. The probabilistic topology-preserving characteristic of GTM has made it a popular tool for data analysis and chemical visualization. Furthermore, it has been used to build successful classification and regression models. The algorithm inserts a two-dimensional "rubber sheet"-like manifold into the initial descriptor space in order to reproduce the best data by a simulated probability distribution function. The latter is represented by an ensemble of Gaussian functions located at the nodes of a two-dimensional grid related to the manifold. Distortion of the manifold is controlled within the limits of a predefined set of parameters. Finally, the molecules are projected with a given probability onto each node of the manifold which then is then projected onto a two-dimensional latent space in which a molecule is associated with one or more nodes.

For each molecule M mapped onto GTM, a probability matrix $R(M, K)$ is calculated which gives the probability of M residing in node K, i.e., the responsibility of node K toward molecule M. Generally, the responsibility of a molecule may be distributed across several nodes. The overall probability to see a molecule anywhere on the map, i.e., $\Sigma_K R(M, K)$ is always equal to 1.0.

The set of compounds S residing in a node K is represented by cumulated responsibilities of K toward all of its members, $\rho(S,K) = \Sigma_{M\in S} R(M, K)$. It represents the density distribution or fuzzy membership of compounds in a set in a particular node of GTM. Here, $\rho(S,K)$ defines the node-bound density distribution of the compound set S.

2.3. GTM Generation and Validation. The manifolds used in this work have been developed and validated previously. In this section, we provide a brief reminder of earlier reported methodology. An in-house ISIDA-GTM[10,18−20] tool was used to build the maps and to develop GTM-based classification models. An evolutionary algorithm-based approach was implemented in order to optimize the parameters required for GTM setup as reported.[2,11] The chromosome vector encoded the current choice of operational parameters. These are a type of ISIDA descriptors,[12−14,21] number of nodes, number of radial basis functions (RBF) defining the manifold, and their width, the regularization coefficient. In addition, a set of compounds called "frame set" is also included in the chromosome, which defines the reference frame for fitting the manifold. Construction of the manifold is done in an unsupervised way, meaning the "frame set" compounds need not be labeled with specific properties. The "frame set" compounds thus only span the relevant chemical space. Several "frame sets" were therefore used to allow the algorithm to span different chemical space zones and select the best quality maps. In order to assess the quality of the maps, a GTM-based regression model was generated for a separate set of compounds called "selection sets" that are labeled with experimental property values (here, pK_i values). For each selection set, the model produces a cross-validated (3-fold) prediction of the property. The predicted values were then compared to their experimental values, and cross-validated determination coefficient Q^2 is obtained. The mean of all set-specific Q^2 finally served as the score to assess the fitness of the map.

Finally, five maps with good fitness scores were sent to the last stage, whereby they were validated using an external set of compounds called "challenge sets". This validation step was carried out to see if new compounds can be well projected onto the generated maps. Two-thirds of the "challenge sets" is used to train the model, i.e., color the map by their property to build the activity color map (ACM). The remaining one-third was then used as a test set, whereby they are classified as active or inactive using the approach reported in ref 18. The process was repeated three times, so that each "challenge set" molecule gets a predicted class value. The final prediction was taken as the one that was obtained in at least two of the three iterations. A cross-validated balanced accuracy was calculated by comparing predicted and actual classes of the compounds in "selection sets". The maps with higher cross-validated balanced accuracy values were able to achieve a larger number of correctly predicted properties and were prioritized as good quality maps. Maps 2, 3, and 5 have been selected for this work.

2.4. "Universal" GTMs for General, Polypharmacologically Competent Mapping of Drug-Like Space. Previously published "universal" GTMs resulted from a quest to develop consensus GTMs with "polypharmacological competence", i.e., the ability to render, on the same map, multiple property landscapes associated with diverse biological activities. The construction of a map—technically, fitting of its manifold ("rubber sheet") in the original descriptor space—is an act of unsupervised machine learning. Manifold construction relies on a "frame set" of compounds (marketed drugs, biological reference compounds, compounds from various databases like ChEMBL, PubChem,[21] and ZINC[22]) chosen such as to span a maximum of the relevant drug-like chemical space zones, without the need to know or specify other experimental properties. Subsequently, supervised generation of property-specific models can be achieved by "coloring" the map by various properties or activities associated with specific compound series. The latter are called "selection sets", which are not mandatorily overlapping with frame sets. Eventually, these "colored" activity[23] or classification[11] landscapes are challenged to serve as predictors for external compounds co-localizing on them.

DOI: 10.1021/acs.jcim.7b00128

The central idea behind the conception of "universal" maps was thus to *select*, using an evolutionary algorithm,[24] those manifolds which passed a maximum of colored landscape-based predictive tests.[11] Selection was done out of a large number of possible manifolds, built in one of 39 eligible ISIDA descriptor spaces[12−14,25] with differently sized node grids.

Finally, the top five maps—corresponding to five descriptor spaces given in Table 1—produced by the evolutionary algorithm were selected and used for predictions.

Table 1. Top Five Universal GTMs and Underlying Descriptor Spaces[a]

map	descriptors	size	#RBF	RBF width
1	pharmacophore-colored atom-centered ISIDA fragment counts, based on sequences of atoms and bonds of fixed length, covering first and second coordination sphere	40	16	1.0
2	CVFF force field type colored ISIDA counts of atom pairs at 1−5 bonds apart, including interposed bond information	32	19	0.9
3	as above, but without bond information	39	17	1.1
4	ISIDA pharmacophore-colored counts of atom pairs found at 1−5 bonds apart, including information on bonds nearest to terminal atoms	32	17	0.6
5	ISIDA pharmacophore triplets, with edges of topological distances 3 and 4	40	15	0.2

[a]Reported are the descriptor space and size of the top five universal maps obtained from an evolutionary optimizer. The size refers to the number of nodes defining the edge length of the square grid representing the GTM. Also reported is the size of the grid locating the radial basis functions (#RBF) and its width.

The predictive challenges mainly involved testing the ability of high scoring GTMs to classify active vs inactive compounds for >400 targets, discriminating ligands based on their preferred target family, and comparing chemical space coverage for related targets, yielding encouraging results. Taken together, the findings suggested that "universal" GTM successfully integrated property distributions with the "geography" of 2D chemical space projections. Further, they are capable of handling large data sets. In an earlier study, ∼1.47 million compounds were mapped onto the universal GTM.[11] The present work mainly aims—following additional external validation tests of the maps—to heuristically analyze the medicinal chemistry "knowledge" behind this *in silico* view of drug-like chemical space.

2.5. GTM-Based Classification Models. A validation aspect not fully addressed previously is the ability of universal maps to discriminate between ligand sets associated with specific target families and other families belonging to the same superfamily. This "family within superfamily" classification is a more challenging task than the previously validated separation of ligand sets for different superfamilies.

Therefore, for each target family within its respective superfamily—according to the ChEMBL ontology—available ligands were grouped together in "family sets". For example, the tyrosine kinase (TK) family set regroups all inhibitors having a reported K_i or IC_{50} value for at least one human tyrosine kinase in ChEMBL. For each superfamily composed of N target families (only families with associated family set sizes of >50 distinct compounds were retained), all $N(N-1)/2$ pairwise interfamily discrimination challenges were performed, using a standard 3-fold cross-validation scheme. The current pair of family sets was merged into a "color" set where members of either family set were given different class labels 1 or 2. Should the same compound be active against targets from both families, it would appear twice in the color set, with both labels, indicating a genuine overlap of the family-specific chemical space zones. Iterative "coloring" of an activity class landscape by two-thirds of the color set was followed by predictive assignment of class labels of the left-out one-third of the color set compounds, with a correct prediction meaning that class separation was effective. Representative families and the number of associated compounds are given in Table 2.

Table 2. Target Family-Based Compound Data Sets[a]

superfamily	target family	abbreviation	#CPDs
proteases	serine proteases	Ser	7585
	metallo proteases	Mmp	4131
	cysteine proteases	Cys	3227
	aspartic proteases	Asp	3068
	threonine proteases	Thr	165
kinases	serine threonine kinases	SerThr	10,804
	tyrosine kinases	Tyr	9907
	PI3/PI4 kinases	PI3/4	1982
GPCRs	short peptide receptors	Sho	14,472
	monoamine receptors	Mon	14,101
	lipid-like ligand receptors	Lip	7613
	nucleotide-like receptors	Nuc	5811
	chemokine receptors	Che	5042

[a]Reported is the composition of compound data sets assembled for three target superfamilies.

Classification landscapes were colored to distinguish regions of the map corresponding to two given classes (red for class "1"; blue for class "2"). The node color of the classification landscape was determined by the class with the majority of compounds assigned to a given node. The class label of a test compound is then extrapolated from the nodes onto which their responsibility distribution will "position" it. Model performance was evaluated by balanced accuracy calculated as follows:

$$BA = \frac{0.5\ TP}{TP + FN} + \frac{0.5\ TN}{TN + FP}$$

where, TP, TN, FP, and FN are true positives, true negatives, false positives, and false negatives, respectively.

2.6. Privileged Responsibility Patterns. Each compound mapped onto a GTM is characterized by its responsibility vector. Compounds with similar responsibility vectors are positioned close to each other in the map, complying with the neighborhood principle,[23] and have similar values of initially calculated chemical descriptors. Responsibility vectors are real-value vectors, and the chance of finding two or more molecules with strictly identical responsibility vectors is very low. A straightforward method to regroup compounds with nearly identical responsibilities is cell-based clustering in responsibility vector space, i.e., tessellating the responsibility vector space into a grid of cubic "cells" and assuming compound members of a same cell being representatives of a common responsibility pattern (RP). Thus, the RP of the compound can be instantly determined by rounding up[26] its responsibility vector. Compounds that have common responsibility patterns are clustered together in the map and provide a basis to further analyze groups of related compounds.

In our analysis, an RP was considered "privileged" by a target (super)family if it characterized a large fraction of compounds belonging to that target (super)family and if its global occurrence rate in the "default" set of compounds was significantly lower.

DOI: 10.1021/acs.jcim.7b00128

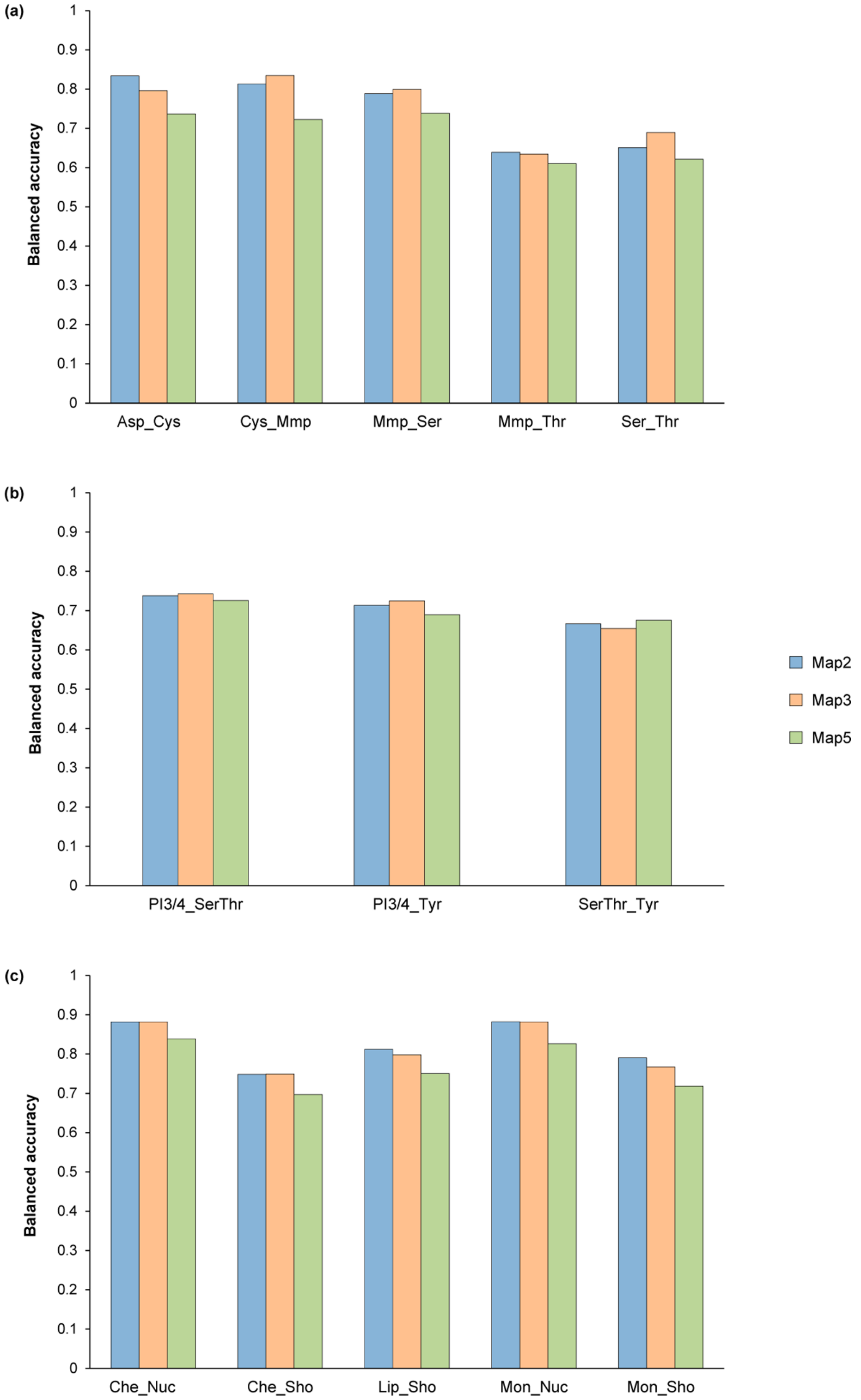


Figure 1. Performance of GTM classification models. The bar chart reports 3-fold cross-validated balanced accuracy of the GTM-based models in classification of (a) protease inhibitors, (b) kinase inhibitors, and (c) GPCR ligands. The performance of maps 2, 3, and 5 are represented by color-coded bars.

Here, the default set serving as the source for these occurrence frequency baselines were the above-mentioned 152,274 (152 K) CHEMBL compounds with high-quality activity data. In this study, RPs privileged toward three superfamilies (protease, GPCR, kinase) were identified following a sequence of steps carried out using the KNIME[27] data analytics tool. The 152 K reference compound set was mapped (using map2 and map3, respectively, Table 1) and compound RPs were extracted. Rarely

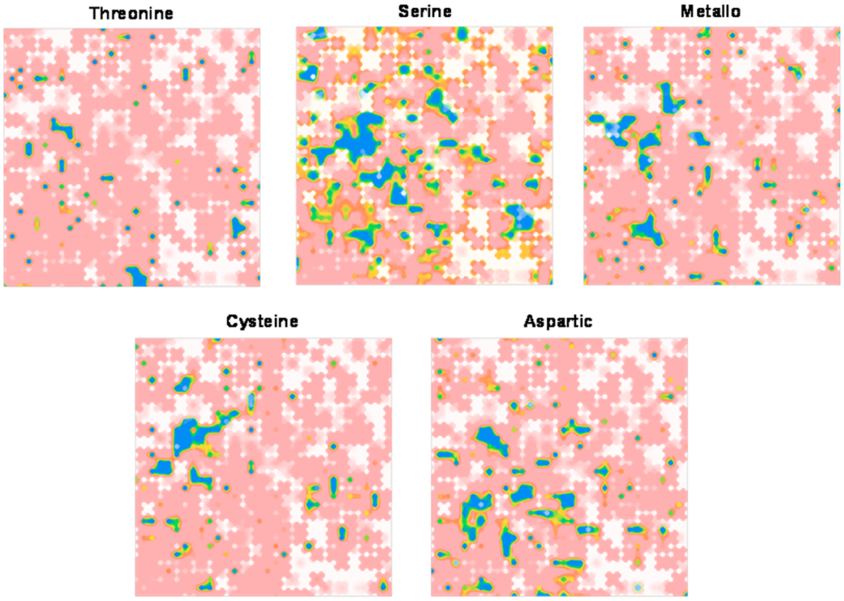


Figure 2. Classification landscapes of protease inhibitors. Shown are the classification landscapes for inhibitors of five protease families. Each map separates regions predominantly occupied by protease inhibitors (blue; class "2") from those populated by other ChEMBL compounds (red; class "1"). The color intensity reflects the density of compounds, and empty nodes are rendered nearly transparent. Relative predominance of compounds is indicated using the following color spectrum: red (class "1"); orange, yellow, green, blue (class "2").

occurring RPs seen in <50 compounds were not considered for further analysis. This resulted in a total of 96,405 and 90,977 compounds represented by 367 and 448 responsibility patterns in map2 and map3, respectively. The fraction of compounds belonging to an RP out of all the compounds mapped onto the GTM was calculated as the global fraction:

$$f\_\text{global (RP)} = \frac{\text{number of compounds matching RP}}{\text{total compounds in map}}$$

Further, a local fraction f_local (RP, SF) was defined as the ratio of compounds active against target superfamily SF that was characterized by RP.

$$f\_\text{local (RP, SF)} = \frac{\text{number of compounds in SF, matching RP}}{\text{total compounds in SF}}$$

Finally, a "privilege" factor was calculated as follows, increasing values of which meaning that RP is privileged by superfamily SF.

$$\text{PF} = f\_\text{local (RP, SF)}/f\_\text{global (RP)}$$

For each target superfamily, the top 10 RPs in the descending order of their privilege factor were designated as PRPs. PF values varied in the range of 3.2−3.4 (GPCRs), 3.7−7.0 (kinases), and 7.1−9.0 (proteases).

2.7. Privileged Structural Motif Detection from PRPs.

Compounds sharing a same map-specific, abstract RP will likely contain some common underlying human-interpretable chemical structural motif. This does not necessarily mean that they share a common scaffold but might have a set of similar scaffolds and perhaps also share some of the substituents around the common scaffold(s). If an RP is privileged by the above-given definition, i.e., if it is a PRP, then its underlying structural motif also inherits the "privileged" status. The process of unveiling the structural motif behind a PRP is thus translating the map-encoded information into plain human knowledge, and ideally, this knowledge shall be complementary to what humans can learn by direct reasoning in terms of substructure. Our search for the common structural motif in each privileged RP-specific compound involved their systematic fragmentation followed by visual inspection. Compounds were subjected to fragmentation based on retrosynthetic (RECAP) rules using an in-house java tool based on the OpenEye[28] chemistry toolkit. The RECAP fragmentation scheme involves systematic cleavage of the bonds of compounds that represent predefined chemical reactions and results in RECAP core and the corresponding substituent. As such, a compound may be fragmented in more than one way (following different RECAP rules) and hence may give rise to more than one core and fragment combination. The number of unique RECAP cores resulting from fragmentation of compounds in a given PRP depends on their structural diversity. Further, a single RECAP core can represent one or multiple compounds of PRP. Any RECAP core that occurred in more than one compound of a given PRP was designated as a "frequent" RECAP core. Visual analysis of the frequent RECAP cores was performed to identify common substructures. Such a substructure was designated a "Privileged Structural Motif" (PSM) shared by a majority of compounds within a given PRP.

3. RESULTS AND DISCUSSION

3.1. Analysis Concept.

Compound locations on the two-dimensional grid of nodes representing a GTM are defined by responsibility vectors. Compounds with similar responsibility vectors are expected to be related. Formally, responsibility vectors were converted into a discretized form called a responsibility pattern (RP), with compounds having the same RP being members of a same cluster in a map.[2] RPs were examined for a "privileged status" with respect to a given target family, and the top 10 representative patterns were privileged responsibility patterns (PRPs). Regions of the map pointing at PRPs were delineated, and corresponding compounds were extracted. Finally, substructures that frequently occurred in

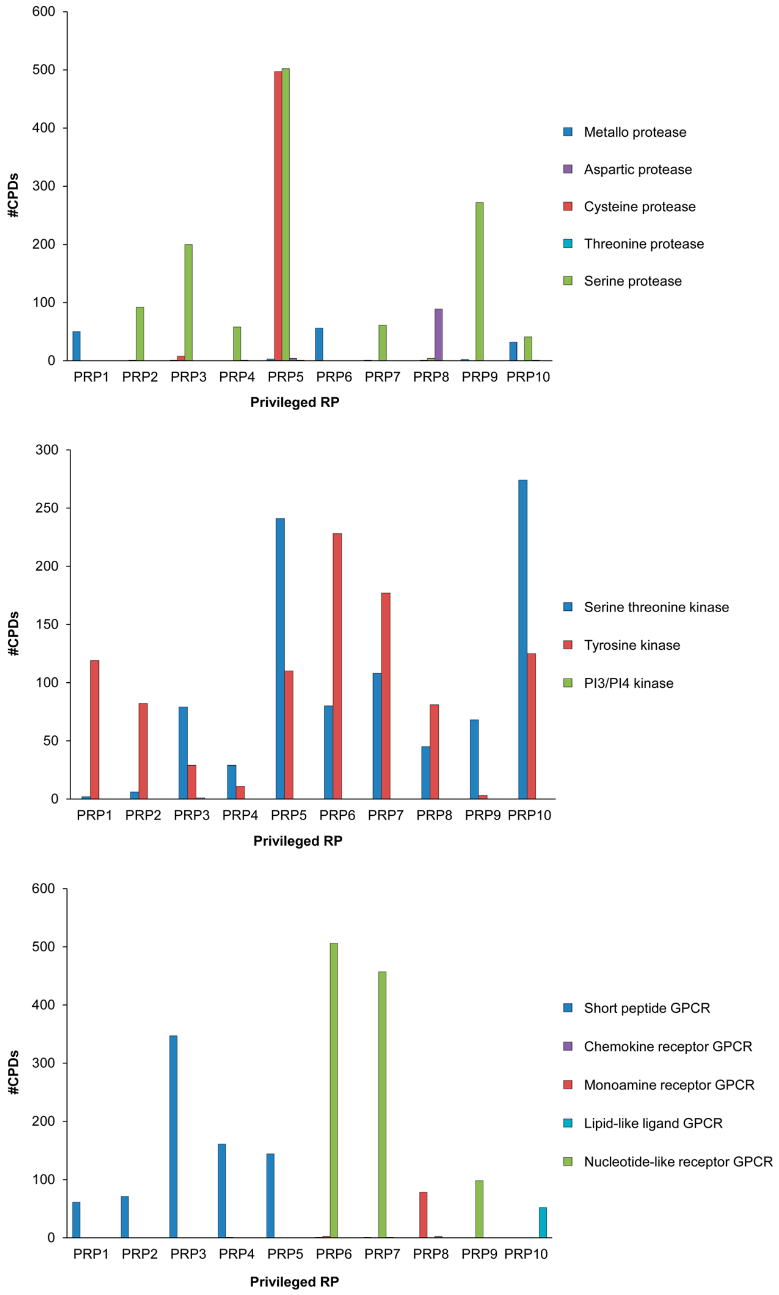


Figure 3. Specificity of PRP compounds toward target families. The bar charts show distribution of compounds containing PRPs for proteases (top), kinases (middle), and GPCRs (bottom). Individual target families are represented by color-coded bars.

molecules representing PRPs were designated as privileged structural motifs if possible and further analyzed.

PRPs self-adaptively focus on the optimal "resolution" needed to define privileged structural motifs. Sometimes, the common signature of all members of a PRP is, indeed, a privileged substructure or scaffold, but more often, it is a set of similar, interchangeable scaffolds or, more general, a pharmacophore pattern compatible with several, distinct carrier scaffolds.

Assessing the privileged status of RPs is a straightforward statistical exercise, but privileged structural motifs must be defined. Unlike in a chemist's viewpoint, however, targets do not adhere to the scaffold-centric view of ligands; rather, they recognize compounds carrying various substituents. Accordingly, going beyond scaffold analysis, it is challenging in compound activity prediction to *a priori* suggest complex structural motifs such as the ones highlighted by GTMs, which are selected on the basis of their propensity to quantitatively explain a broad spectrum of structure—activity data. Herein, identified PRPs and the associated privileged structural motifs focus the analysis on the most relevant structural patterns, going a step further than it is possible based upon the facile, scaffold-centric view of the structure—activity relationships.

3.2. Performance of Target Family Classification Models.

Figure 1 shows the classification performance reported as cross-validated balanced accuracy for three universal GTM. The settings used for each map are given in Table 1. Inhibitors of

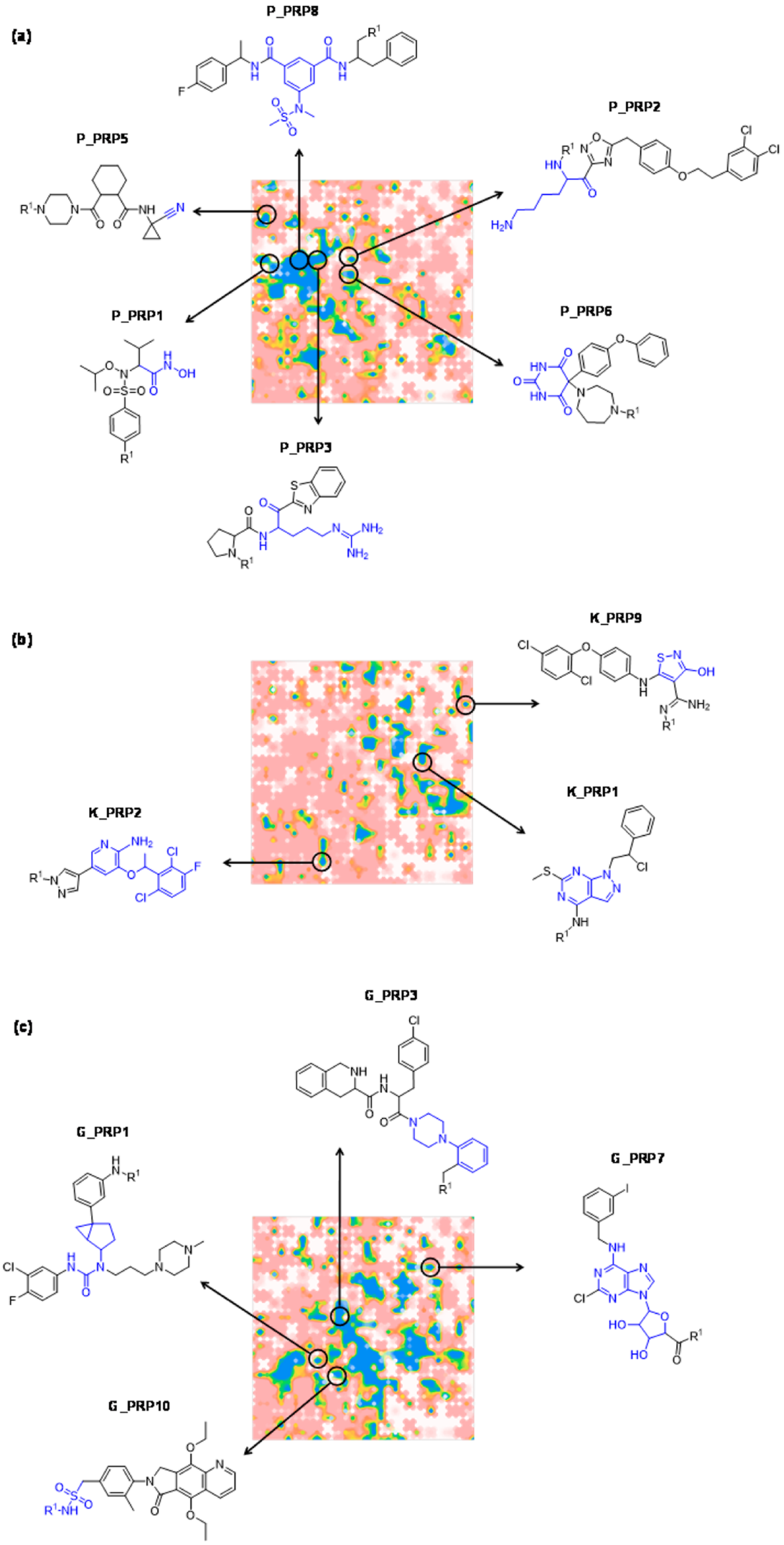


Figure 4. Privileged structural motifs. Given are the regions of GTM (map2) where compounds containing privileged structural motifs of (a) protease inhibitors, (b) kinase inhibitors, and (c) GPCRs are localized. In each case, a frequent RECAP core is shown that represents an individual PRP and contains a privileged structural motif highlighted in blue.

different protease families were overall well separated with 3-fold cross-validated balanced accuracy ranging from 0.6 to 0.85. These findings matched expectations. Better than random separation levels were mandatory, whereas perfect separation of target family ligands is not a realistic expectation since cross-family ligands are available. A partial overlap of family-specific chemical space zones may be a natural consequence of cross-family "promiscuity" and not necessarily a classification model failure. Separation performance displayed only limited data set dependence. Threonine protease inhibitors were an exception because they were poorly predicted across all three maps. This was due to the small size of this data set (comprising only 165 compounds) compared to the other protease families (for which the sets were an order of magnitude larger). The problem was due to compound set size imbalance in the context of aggressive 3-fold cross-validation: when the set of threonine protease inhibitors was merged as class 1 members into a "color set" where the number of class 2 members belonging to the other target families was much larger, random splitting into two-thirds vs one-third led to subsets that were perfectly balanced with respect to class 2 compounds but unbalanced with respect to the much smaller number of class 1 examples. Without cross-validation, the threonine protease inhibitors occupied a well-defined series of activity "islands" on the map within the background of the entire ChEMBL database (Figure 2).

Classification of serine/threonine vs tyrosine kinase inhibitors was overall more challenging than the other two kinase family pairings, with cross-validated balanced accuracy of 0.67. By contrast, GPCR classification models reached balanced accuracy scores as high as 0.88 and had an overall cross-validated balanced accuracy of at least 0.75. Map2 and map3 performed equally well throughout most of the classification trials and outperformed map5. Accordingly, fragments annotated with force field scores were preferred as descriptors over pharmacophore triplets for classification of inhibitors and ligands of the three target superfamilies. The results from classification models clearly indicate that the universal maps used here were able to separate regions of chemical space populated by compounds specific for individual target families.

Figure 2 shows classification landscapes for the five protease families obtained from map2. Each map distinguishes the regions of the landscape that are densely populated by compounds active against the respective protease family (blue; class "2") in contrast to the rest of the ChEMBL compounds (red; class "1").

The patches of yellow/green indicate the areas where the two classes overlap—meaning that ligands associated with a given target family have very near neighbors in ChEMBL *not* associated with the same family. In this context, "associated" means that a compound has been tested against at least one family member in dose−response experiments. Therefore, there are possible implications. The existence of nonassociated neighbors provides an opportunity for virtual screening, unless these compounds were found to be inactive in primary screens or at least not interesting enough to justify a dose−response measurement follow-up.

The presence of several dark blue zones shows that protease-specific compounds were clustered together in the map, visually illustrating the good predictive power of classification models. As expected, positions of blue regions in each of the five maps were distinct and indicated the presence of target family-specific clusters of compounds. Serine proteases occupied a larger area of the landscape compared to other family-directed compounds because they represented the largest group and covered more than 40% of all protease inhibitors.

3.3. Responsibility Patterns Privileged by Target Superfamilies.

The following reports the analysis of the RPs selected for their high privilege scores with respect to either of the three superfamilies—proteases, kinases, and GPCRs. Albeit the RPs were, as mentioned, picked for their superfamily-related privilege scores, further analysis will show that they often "spontaneously" regroup ligands of specific target families within the large superfamily—actually, an expected result, with retrospect to the robust target family separation scores obtained in quantitative validation, above.

Compound and target family distributions among PRPs of protease, GPCR, and kinase inhibitors are monitored in Figure 3. The bar plots for proteases indicate uneven distribution of PRPs across inhibitors, ranging from 50 compounds for PRP1 to 1007 for PRP5. It was an interesting observation that the majority of compounds for a given PRP were active against a single target family, with few exceptions such as PRP5 and PRP10. Further, PRPs of kinase inhibitors were dominated by compounds belonging to one of the three kinase families. Actually, more than 99% of the kinase inhibitors displayed activity against some representative of these three target families, which explains these observations. The dominance of a single family for individual PRPs was not the rule for the kinase superfamily, in contrast to proteases where patterns were mostly family specific and GPCRs where all patterns were family specific. All GPCR ligands with PRP1 to PRP5 targeted the short peptide GPCR family, whereas the majority of ligands for PRP6 and PRP7 were active against nucleotide-like receptor GPCRs. For GPCRs and proteases, the maps tended to show the existence of a privileged superfamily-specific chemical subspace emerging as the union of an essentially disjoint family-specific PRPs.

The kinase superfamily also defined a dedicated chemical space zone, which was, however, less clearly subdivided into family PRPs. This might result from the fact that most—but not all—currently available kinase inhibitors are directed against the ATP cofactor binding site, which is largely conserved across the Tyr and Ser/Thr kinase families (and also bears resemblance to ATP binding sites in other proteins). As a consequence, subsets of ATP site-directed kinase inhibitors tend to be promiscuous and frequently inhibit multiple kinases. Out of all the kinase inhibitors included in our study, 9.2% inhibited at least three targets and 73.5% were active against single kinase. The finding that PRPs displayed specificity toward particular target families provided corroborating evidence that grouping of active compounds on GTM based on their RPs was meaningful. Compounds associated with an PRP are likely to have similar features and therefore would most likely bind to the same or related targets.

3.4. Analysis of Extracted Privileged Structural Motifs.

The approach to subject active compounds to RECAP fragmentation and subsequently visually analyze frequent RECAP cores led to identification of several structural motifs linked to privileged patterns. Figure 4 shows RECAP cores containing privileged structural motifs mapped onto GTM (map2) from which the PRPs were extracted. The RECAP cores shown in Figure 4 are representative examples of most frequent cores extracted from compounds associated with given PRPs, with the common, strictly conserved part, the "privileged structural motif" (PSM) being highlighted in blue. In clear, this means that compounds within the given PRP all strictly conserve

DOI: 10.1021/acs.jcim.7b00128

a

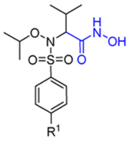
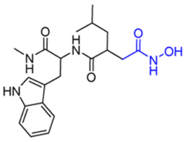
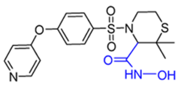
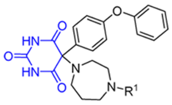
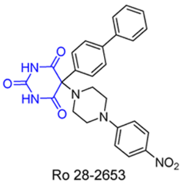
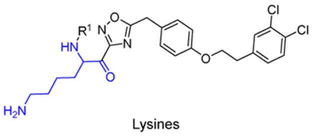
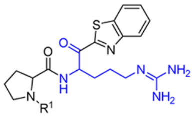
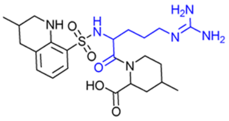
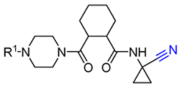
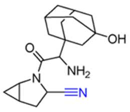
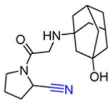
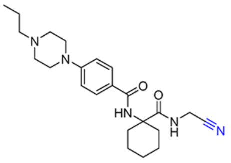
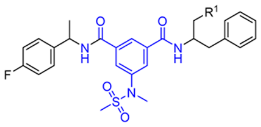
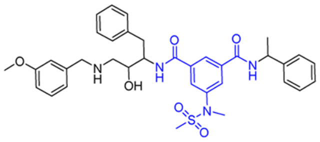
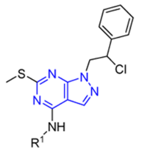
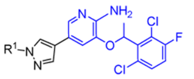
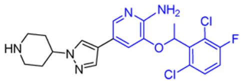
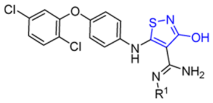
PRP	Occurence	Frequent RECAP core	Representative compounds of PSM
P_PRP1	100%	 Hydroxamates	 Ilomastat  Prinomastat
P_PRP6	100%	 Barbituric acids	 Ro 28-2653
P_PRP2	100%	 Lysines	
P_PRP3	23.9%	 Arginines	 Argatroban
P_PRP5	94.9%	 Carbonitriles	 Saxagliptin *  Vildagliptin  Balicatib *
P_PRP8	73.4%	 Isophthalimides	 GRL-8234

Figure 5. continued

b

PRP	Occurence	Frequent RECAP core	Representative compounds of PSM
K_PRP1	99.2%	 Pyrazolo[3,4-*d*]pyrimidines	
K_PRP2	59.8%	 2-Amino-5-aryl-3-benzyloxypyridines	 Crizotinib *
K_PRP9	67.1%	 Isothiazoles	

c

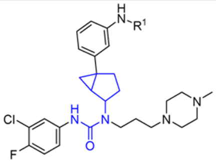
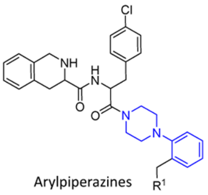
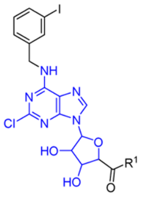
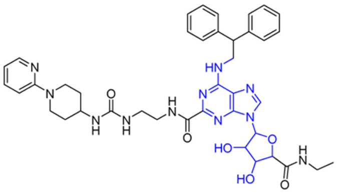
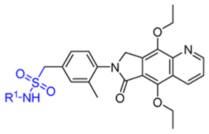
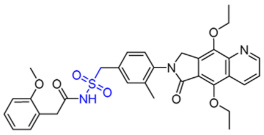
PRP	Occurence	Frequent RECAP core	Representative compounds of PSM
G_PRP1	67.2%	 Bicyclo[3.1.0]hexyl ureas	
G_PRP3	79%	 Arylpiperazines	
G_PRP7	49.9%	 Adenosine derivatives	 Uk-432097 *
G_PRP10	57.7%	 Sulfonamides	 MF498

Figure 5. (a) Privileged structural motifs of protease inhibitors. (b) Privileged structural motifs of kinase inhibitors. (c) Privileged structural motifs of GPCR ligands. Shown are the most frequent RECAP cores for each PRP of protease inhibitors with the associated PSM highlighted in blue next to exemplary compounds illustrating the relevance of the PSM for inhibition. These exemplary compounds include classical types of inhibitors and drugs. The last column is left blank if no such exemplary compound containing a given PSM was selected and discussed in the text. If the reference compounds

Figure 5. continued

are actually members of the PRP, they are highlighted by an asterisk—otherwise, they map to some other RP supporting the current PSM or they were not within the data set analyzed herein.

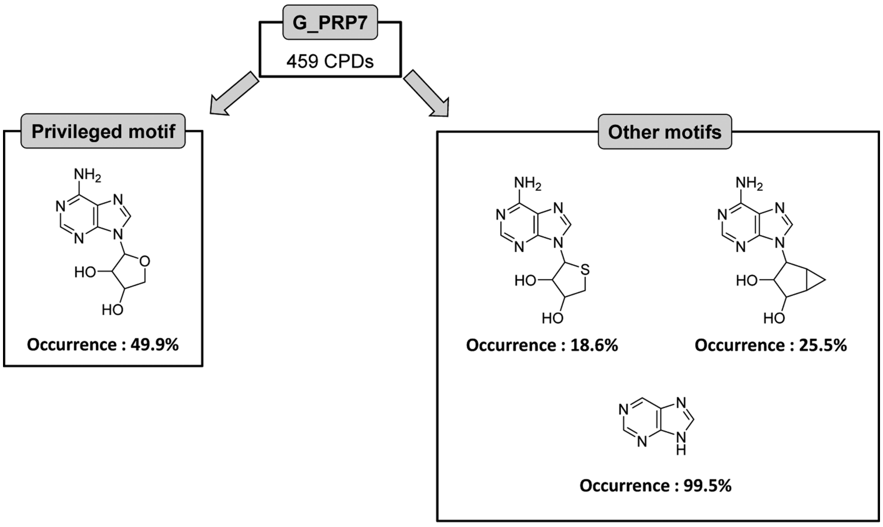


Figure 6. Privileged and other common structural motifs for PRP7 of GPCR ligands. The privileged structural motif and three other common motifs derived from G_PRP7 are given with their frequency of occurrence in compounds with G_PRP7 membership.

the highlighted privileged structural motif and are otherwise structurally very similar to the nonhighlighted moieties.

Figure 5 shows for each highlighted PSM exemplary compounds from the medicinal chemistry literature, if available. These compounds include "classical" inhibitors that are characteristic for specific target and current drugs. Thus, in these cases, compounds of high medicinal chemistry relevance were available that contained signature motifs identified by responsibility patters. We emphasize that many active compounds contain the core identified in our analysis. However, the exemplary compounds shown in Figure 5 represent selected "highlights" from the practice of medicinal chemistry that further support the potential of GTM-based responsibility pattern analysis. We also note that same PSM may occur in more than one responsibility pattern (privileged or not); therefore, the "reference" compounds picked from the literature may, but *must* not, be a representative of the PRP used to highlight the PSM. When this is the case, the reference compound will be marked by an asterisk. Frequent RECAP cores for each target family PRP (designated as P_PRP, G_PRP, and K_PRP) contained structural motifs known to often occur in bioactive compounds. The frequent RECAP cores for the top 10 PRPs of each family are provided in Table S1 of the Supporting Information.

Except for P_PRP3, all highlighted structural motifs from protease PRPs represented more than 70% of the protease inhibitors for the respective PRP. P_PRP 1 and 6 characterized known protease inhibitor types acting as chelating agents for the catalytic Zn^{2+} ion in the active site of metalloproteinases.[29] Hydroxamate-based inhibitors such as ilomastat[30] and prinomastat[31] were among the first generation of matrix metalloproteinase inhibitors, whereas Ro 28-2653 exemplified an inhibitor class based on the barbituric acid scaffold.[32] Furthermore, RECAP cores of P_PRP 2 and 3 included lysine and arginine residues and heterocyclic-substituted ketones, providing an activated warhead for irreversible protease inhibition. These basic aliphatic motifs are highly common among inhibitors of thrombin-like serine proteases, for example, in argatroban, an approved competitive thrombin inhibitor.[33] Thrombin-like enzymes display a primary substrate specificity for basic amino acids in the P1 position, i.e., the amino acid N-terminal of the scissile peptide bond. The corresponding residues such as arginine are recognition elements for the interaction with the S1 binding pocket.[34]

P_PRP5 included a variety of peptide nitriles. Serine and cysteine proteases are predestinated to be inhibited through a covalent bond with the inhibitor because these proteases operate via a covalent mode of catalysis. Accordingly, peptidomimetic compounds with a carbonitrile moiety in place of the scissile peptide bond were widely investigated as serine or cysteine protease inhibitors and shown to form imidates and thioimidates, respectively.[35] Saxagliptin[36] and vildagliptin, approved antidiabetic drugs, are representative examples of dipeptidyl peptidase-4 inhibitors.[37] Balicatib was developed as a potent inhibitor of the cysteine protease cathepsin K[38] but failed in clinical trials. Isophthalamides were represented in PRP8 and known to be inhibitors of memapsin 2,[39] an aspartic protease involved in Alzheimer's disease. GRL-8234 displays an exemplary three-winged isophthalamide with its three substituents interacting specifically with the S3, S2, and S1 pockets of memapsin 2.[40]

Furthermore, PRPs of kinases were also associated with typical inhibitor scaffolds. For K_PRP1, a class of pyrazolo[3,4-*d*]pyrimidine derivatives exemplified dual inhibitors for Src and Abl tyrosine kinases.[41] Co-crystallization of pyrazolopyrimidines bound to protein kinases indicated a mode of action, which was based on a specific disruption of a glutamine−lysine interaction, resulting in an inactive conformation of the kinases.[42] The highlighted structural motif was contained in 99% of molecules defined by K_PRP1. K_PRP2 consists of halogenated 2-amino-5-aryl-3-benzyloxypyridines. Kinases are activated by the phosphorylation of the activation loop, which acts as an autoinhibitor in its dephosphorylated state.[43] The potency and selectivity of crizotinib, approved for the treatment of nonsmall cell lung carcinoma, results from key interactions within the lipophilic pocket of its target. The (*R*)-methylbenzyloxy moiety of crizotinib binds to the activation loop and stabilizes the dephosphorylated state. Thus, the kinase remains inactive.[44] Isothiazoles, as represented in K_PRP9, have been investigated toward their inhibitory potential on MEK1, a serine/threonine specific protein kinase.[45]

Moreover, the strategy of modifying natural peptide hormones as inhibitors of GPCRs was exemplified by the RECAP cores associated with of G_PRP1 and 3.[46] A bicyclo[3.1.0]hexyl urea series was reported for its inhibitory potential against melanin-concentrating hormone receptor-1,[47] whereas arylpiperazines with a dipeptide substitution showed inhibitory activity against melanocortin-4 receptors.[48] The bicyclic adenine core and the ribose ring, presented in G_PRP7, are key features of adenosine receptor agonists. UK-432097 was developed as a highly selective adenosine A2A receptor agonist and was later discontinued from clinical trials.[49] G_PRP10 was associated with fragments of sulfonamides connected to a tricyclic (aza)naphthalene scaffold. The acylsulfonamide MF498 is an example for an E-type prostanoid receptor 4 inhibitor.[50]

The structural motifs that are discussed above were found to be most ubiquitous among individual PRP members. As mentioned earlier, they might be found in all or only a subset of PRP-represented molecules. The frequency of their occurrence given in Figure 5 ranged from 38.8% (P_PRP8) to as high as 100% (P_PRP1, 2, and 6). On a closer look, it was observed that the compounds that lacked the highlighted structural motifs often contained very similar motifs. Figure 6 lists other commonly seen structural motifs of G_PRP7 compounds that were either substructures of the identified privileged motif or had minor structural differences.

Although the privileged structural motif represented nearly half of the compounds in G_PRP7, it is evident that replacement of oxygen by a sulfur atom in the ribose ring of this motif gives rise to another motif representing 18.6% of the compounds. Furthermore, another new motif obtained by entirely replacing the ribose ring with bicyclo[3,1,0]hexane accounts for 25.5% of all the compounds.

3.5. Evolution of Privileged Patterns in ChEMBL.

As demonstrated, privileged structural motifs identified by our methodology were frequently reported in medicinal chemistry literature to be specific for an inhibitor type and, in a number of instances, were also found in approved drugs. We reasoned that it would also be interesting to examine how these structural motifs might have evolved in ChEMBL. This would be helpful to determine at what stage of chemical exploration inhibitors containing these specific structural motifs occurred. Therefore, for each PRP associated with a privileged structural motif, the occurrence of corresponding compounds in ChEMBL was monitored on a time course (Figure 7). It is evident that P_PRP5 experienced much interest in protease inhibitor research because the majority of inhibitors reported over the past decade included P_PRP5-associated motifs. Also, most of the compounds with P_PRP2 and 3 were reported in single year (2006 for P_PRP2) or two years (2005 and 2008 for P_PRP8), but their popularity decreased in subsequent years.

Although kinase research dates back to the early 1980s, rapid exploration of kinase inhibitors occurred more recently beginning in 2004. Interestingly, compounds with isothiazole, as represented by K_PRP9, which are the most extensively explored kinase inhibitors, were reported in 2006. In case of GPCRs, compounds associated with individual PRPs were continuously explored and reported over time. GPCR ligands containing arylpiperazine (G_PRP3) and adenosine derivatives (G_PRP7) have been particularly popular candidates for compound design over the past few years.

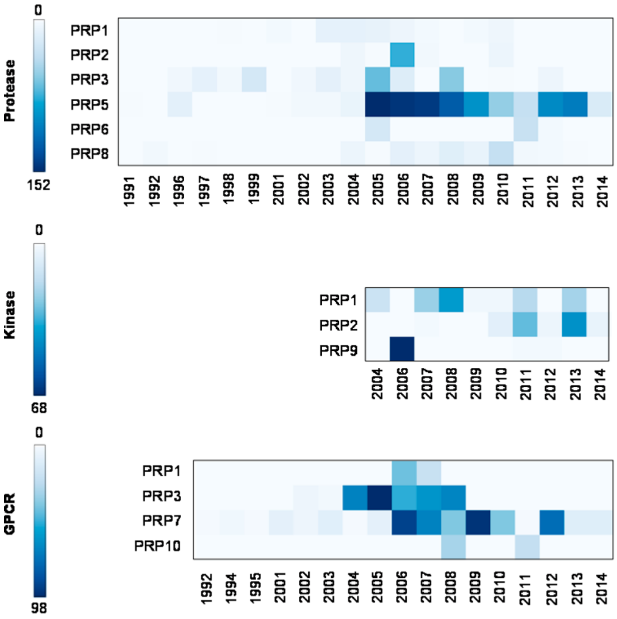


Figure 7. Evolution of PRPs in ChEMBL. Shown are heat maps representing the frequency of compounds containing privileged structural motifs monitored in ChEMBL over time. Cells are color-coded according to compound numbers using a spectrum from white to light blue to dark blue.

4. CONCLUSIONS

Although many studies have been carried out in the past to search for and describe privileged structural motifs, they were mostly scaffold-centric. We have addressed the task of target-family specific pattern detection in molecules using universal GTMs, which currently are among the polypharmacologically most competent 2D mapping representations of drug-like space. In the present work, accurate target family-based classification of inhibitors and ligands of three major super families (protease, kinase, and GPCR) was carried out to further externally validate the use of universal GTMs. Classification tasks consistently returned cross-validated balanced accuracy (BA) levels well above random classification, with the exception of a single small data set for which cross-validation led to substantial size imbalance.

Compound location on GTMs was defined on the basis of their responsibility vectors, and thus compounds with similar responsibility vectors were expected to be related. Responsibility patterns, a discretized from of responsibility vectors, can be used for rapid clustering of related compounds. A frequency analysis-based prioritization of responsibility patterns was done in the quest for PRPs that were much more frequent within actives of a given target superfamily than within the entire compound background universe.

RECAP fragmentation permitted the identification of frequent cores, and the subsequent medicinal chemistry analysis of individual PRP compounds led to the detection of privileged structural motifs that are very often characteristic of target family directed compounds, thus lending credence to the medicinal chemistry relevance of RP and PRP exploration on the basis of GTMs. Such motifs were reported in medicinal chemistry literature to be characteristic features of active compounds and even approved drugs. The results have shown that the GTM-based approach was suitable to prioritize compound clusters that were indeed representative of a target family. Universal GTMs might thus be considered to represent a reasonably educated "expert system" for medicinal chemistry capable of generating a detailed view of key compound classes, without any *a priori* focus on the nature of privileged structural motifs. Although a single RECAP core was often sufficient to detail the structural motif associated with a given PRP, patterns reconciling several related RECAP cores were also detected. This might be particularly useful when analyzing very large compound collections where visual inspection of individual molecules is nontrivial. Furthermore, we have shown that grouping of compounds based on RPs and PRPs goes beyond classical scaffold-based clustering because GTMs account for relationships between similar scaffolds. In cases where PRPs matched multiple scaffolds, they were typically distinguished by only small chemical modifications. Taken together, these findings demonstrate that GTMs accounted for a spectrum of patterns, scaffolds, and privileged structural motifs that were characteristic of compounds active against individual target families.

Two key benefits of the GTM-based view of privileged "patterns" could be delineated:

1. Out of the many possible compound clustering approaches—which often tend to produce highly diverging clustering schemes—this is one of the rare which can be objectively defended on the basis of the excellent behavior of the underlying maps in quantitative classification and regression models. The maps used here have been extensively shown to support predictive, robust activity and class "landscapes" with respect to a plethora of different bioactivities and target class associations—both in previous work and in the present article. Therefore, compound clusters as presented by responsibility patterns are not just "some" way to regroup compounds together, given some descriptor space, dissimilarity metric, and clustering algorithm. They are produced by a mapping scheme evolved in order to be, and quantitatively proven to be, neighborhood behavior-compliant, which is rarely the case in clustering exercises. Furthermore, as GTMs are "fuzzy" counterparts of Kohonen maps, of tunable fuzziness, they are intrinsically able to mimic any results that might have been obtained by this "classical" and widely accepted approach.
2. No human preconceptions are at the basis of the highlighted patterns. Checking the "privileged" status of a structural motif is, *per se*, a trivial task—in as far as, however, a human agent *considers* checking the status of that particular pattern. That is the key reason for the popularity of the privileged scaffold approach: there is a—relatively—limited set of scaffolds to browse through and to check, one by one, the privileged status of each (provided there is agreement upon the definition of the "scaffold" concept, for which several competing paradigms coexist). Browsing through all the imaginable *structural* patterns, e.g., sequentially asking all possible questions, such as "is 1-aminoalkyl-substituted benzindole/imidazole a privileged motif?" is practically not feasible. Yet, as in the example above, the key structural motif causally related to activity will most likely transcend the artificial scaffold/substituent distinction, or the GTM-driven clustering has the merit to *propose* key motifs out of the virtual infinity of possibilities. RECAP analysis has been used in order to inspect the content of the RP clusters according to the conventional substructure-based medicinal chemistry viewpoint and showed that these include, as already discussed, both scaffold and substituent moieties or regroup similar scaffolds into a same pattern if relevant. It is not claimed herein that the highlighted patterns need to be mechanistically relevant, but their relevance reposes on the proven neighborhood behavior compliance, as mentioned above.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jcim.7b00128.

Table S1 reports frequent RECAP cores for the top 10 privileged responsibility patterns of protease, kinase, and GPCR. Raw data (compound SMILES, molecular descriptor files) are available upon request. (PDF)

■ AUTHOR INFORMATION

Corresponding Authors

*E-mail: bajorath@bit.uni-bonn.de (J.B.).
*E-mail: varnek@unistra.fr (A.V.).

ORCID

Dragos Horvath: 0000-0003-0173-5714
Jürgen Bajorath: 0000-0002-0557-5714
Alexandre Varnek: 0000-0003-1886-925X

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

S.K. thanks the Region Alsace for a Ph.D. fellowship. We are grateful to OpenEye Scientific Software, Inc., for the free academic license of the OpenEye Toolkits.

■ ABBREVIATIONS:

RP = Responsibility Pattern; PRP = Privileged Responsibility Pattern; RECAP = Retrosynthetic Combinatorial Analysis Procedure; GTM = Generative Topographic Map; SAR = Structure−Activity Relationship; PSM = Privileged Structural Motif

■ REFERENCES

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

(2) Klimenko, K.; Marcou, G.; Horvath, D.; Varnek, A. Chemical Space Mapping and Structure-Activity Analysis of the ChEMBL Antiviral Compound Set. *J. Chem. Inf. Model.* **2016**, *56*, 1438−1454.

(3) Rokach, L.; Maimon, O. *Data Mining with Decision Trees: Theory and Applications*; World Scientific: Singapore, 2014.

(4) 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 1. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 1017−1026.

(5) Young, S. S.; Hawkins, D. M. Analysis of a 29 Full Factorial Chemical Library. *J. Med. Chem.* **1995**, *38*, 2784−2788.

(6) Bishop, C. M.; Svensén, M.; Williams, C. K. GTM: The Generative Topographic Mapping. *Neural Comput.* **1998**, *10*, 215−234.

(7) Hearst, M. A.; Dumais, S. T.; Osuna, E.; Platt, J.; Scholkopf, B. Support Vector Machines. *IEEE Intell. Syst. their Appl.* **1998**, *13*, 18−28.

(8) John, G. H.; Langley, P. Estimating Continuous Distributions in Bayesian Classifiers. In *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, Montreal, Quebec, Canada, August 18−20, 1995, pp 338−345.

(9) Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5−32.

(10) Gaspar, H. A.; Baskin, I. I.; Marcou, G.; Horvath, D.; Varnek, A. Stargate GTM: Bridging Descriptor and Activity Spaces. *J. Chem. Inf. Model.* **2015**, *55*, 2403−2410.

(11) 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.

(12) Ruggiu, F.; Marcou, G.; Varnek, A.; Horvath, D. ISIDA Property-Labelled Fragment Descriptors. *Mol. Inf.* **2010**, *29*, 855−868.

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

(14) Varnek, A.; Fourches, D.; Solov'Ev, V.; Klimchuk, O.; Ouadi, A.; Billard, I. Successful ″in Silico″ Design of New Efficient Uranyl Binders. *Solvent Extr. Ion Exch.* **2007**, *25*, 433−462.

(15) Strasbourg Virtual Screening Web Server. http://infochim.u-strasbg.fr/webserv/VSEngine.html (accessed July 1, 2016).

(16) Gaulton, A.; Bellis, L. J.; Bento, A. P.; Chambers, J.; Davies, M.; Hersey, A.; Overington, J. P.; et al. ChEMBL: A Large-Scale Bioactivity Database for Drug Discovery. *Nucleic Acids Res.* **2012**, *40*, D1100−D1107.

(17) Wu, C. H.; Apweiler, R.; Bairoch, A.; Natale, D. A.; Barker, W. C.; Boeckmann, B.; Magrane, M. The Universal Protein Resource (UniProt): An Expanding Universe of Protein Information. *Nucleic Acids Res.* **2006**, *34*, D187−D191.

(18) Gaspar, H. A.; Marcou, G.; Horvath, D.; Arault, A.; Lozano, S.; Vayer, P.; Varnek, A. Generative topographic mapping-based classification models and their applicability domain: application to the biopharmaceutics drug disposition classification system (BDDCS). *J. Chem. Inf. Model.* **2013**, *53* (12), 3318−3325.

(19) 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. Inf.* **2012**, *31*, 301−312.

(20) ISIDA Fragmentor software, Laboratoire de Chémoinformatique, University of Strasbourg, 2016. http://infochim.u-strasbg.fr/spip.php?rubrique41 (accessed April 2017).

(21) Wang, Y.; Suzek, T.; Zhang, J.; Wang, J.; He, S.; Cheng, T.; Shoemaker, B. A.; Gindulyte, A.; Bryant, S. H. PubChem Bioassay: 2014 Update. *Nucleic Acids Res.* **2014**, *42*, D1075−D1082.

(22) Irwin, J. J.; Shoichet, B. K. ZINC-a Free Database of Commercially Available Compounds for Virtual Screening. *J. Chem. Inf. Model.* **2005**, *45*, 177−182.

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

(24) Horvath, D.; Brown, J. B.; Marcou, G.; Varnek, A. An Evolutionary Optimizer of Libsvm Models. *Challenges* **2014**, *5*, 450−472.

(25) Nomenclature of ISIDA fragments; Laboratoire de Chemoinformatique, University of Strasbourg, 2012.

(26) Horvath, D.; Koch, C.; Schneider, G.; Marcou, G.; Varnek, A. Local Neighborhood Behavior in a Combinatorial Library Context. *J. Comput.-Aided Mol. Des.* **2011**, *25*, 237−252.

(27) Berthold, M. R.; Cebron, N.; Dill, F.; Gabriel, T. R.; Kötter, T.; Meinl, T.; Ohl, P.; Thiel, K.; Wiswedel, B. KNIME-the Konstanz Information Miner: Version 2.0 and Beyond. *AcM SIGKDD Explor. Newsl.* **2009**, *11*, 26−31.

(28) *OpenEye Scientific Software*; Santa Fe, NM, 2012.

(29) Vandenbroucke, R. E.; Libert, C. Is There New Hope for Therapeutic Matrix Metalloproteinase Inhibition? *Nat. Rev. Drug Discovery* **2014**, *13*, 904−927.

(30) Balakrishnan, A.; Patel, B.; Sieber, A. S.; Chen, D.; Pachikara, N.; Zhong, G.; Cravatt, B. F.; Fan, H. Metalloprotease Inhibitors GM6001 and TAPI-0 Inhibit the Obligate Intracellular Human Pathogen Chlamydia Trachomatis by Targeting Peptide Deformylase of the Bacterium. *J. Biol. Chem.* **2006**, *281*, 16691−16699.

(31) Bissett, D.; O'Byrne, K. J.; von Pawel, J.; Gatzemeier, U.; Price, A.; Nicolson, M.; Mercier, R.; Mazabel, E.; Penning, C.; Zhang, M. H.; Collier, M. A.; Shepherd, F. A. Phase III Study of Matrix Metalloproteinase Inhibitor Prinomastat in Non-Small-Cell Lung Cancer. *J. Clin. Oncol.* **2005**, *23*, 842−849.

(32) Lein, M.; Jung, K.; Ortel, B.; Stephan, C.; Rothaug, W.; Juchem, R.; Johannsen, M.; Deger, S.; Schnorr, D.; Loening, S.; et al. The New Synthetic Matrix Metalloproteinase Inhibitor (Roche 28−2653) Reduces Tumor Growth and Prolongs Survival in a Prostate Cancer Standard Rat Model. *Oncogene* **2002**, *21* (13), 2089−2096.

(33) Chirgadze, N. Y.; Sall, D. J.; Briggs, S. L.; Clawson, D. K.; Zhang, M.; Smith, G. F.; Schevitz, R. W. The Crystal Structures of Human α-Thrombin Complexed with Active Site-Directed Diamino Benzo [B] Thiophene Derivatives: A Binding Mode for a Structurally Novel Class of Inhibitors. *Protein Sci.* **2000**, *9*, 29−36.

(34) Barré, O.; Dufour, A.; Eckhard, U.; Kappelhoff, R.; Béliveau, F.; Leduc, R.; Overall, C. M. Cleavage Specificity Analysis of Six Type II Transmembrane Serine Proteases (TTSPs) Using PICS with Proteome-Derived Peptide Libraries. *PLoS One* **2014**, *9*, e105984.

(35) Fleming, F. F.; Yao, L.; Ravikumar, P. C.; Funk, L.; Shook, B. C. Nitrile-Containing Pharmaceuticals: Efficacious Roles of the Nitrile Pharmacophore. *J. Med. Chem.* **2010**, *53*, 7902−7917.

(36) Metzler, W. J.; Yanchunas, J.; Weigelt, C.; Kish, K.; Klei, H. E.; Xie, D.; Zhang, Y.; Corbett, M.; Tamura, J. K.; He, B.; Hamann, L. G.; Kirby, M. S.; Marcinkeviciene, J. Involvement of DPP-IV Catalytic Residues in Enzyme-Saxagliptin Complex Formation. *Protein Sci.* **2008**, *17*, 240−250.

(37) Juillerat-Jeanneret, L. Dipeptidyl Peptidase IV and Its Inhibitors: Therapeutics for Type 2 Diabetes and What Else? *J. Med. Chem.* **2013**, *57*, 2197−2212.

(38) Palmer, J. T.; Bryant, C.; Wang, D. X.; Davis, D. E.; Setti, E. L.; Rydzewski, R. M.; Venkatraman, S.; Tian, Z. Q.; Burrill, L. C.; Mendonca, R. V.; Springman, E.; McCarter, J.; Chung, T.; Cheung, H.; Janc, J. W.; McGrath, M.; Somoza, J. R.; Enriquez, P.; Yu, Z. W.; Strickley, R. M.; Liu, L.; Venuti, M. C.; Percival, M. D.; Falgueyret, J. P.; Prasit, P.; Oballa, R.; Riendeau, D.; Young, R. N.; Wesolowski, G.; Rodan, S. B.; Johnson, C.; Kimmel, D. B.; Rodan, G. D. And Synthesis of Tri-Ring P3 Benzamide-Containing Aminonitriles as Potent, Selective, Orally Effective Inhibitors of Cathepsin K. *J. Med. Chem.* **2005**, *48*, 7520−7534.

(39) Stachel, S. J.; Coburn, C. A.; Steele, T. G.; Jones, K. G.; Loutzenhiser, E. F.; Gregro, A. R.; Rajapakse, H. A.; Lai, M. T.; Crouthamel, M. C.; Xu, M.; Tugusheva, K.; Lineberger, J. E.; Pietrak, B. L.; Espeseth, A. S.; Shi, X. P.; Chen-Dodson, E.; Holloway, M. K.; Munshi, S.; Simon, A. J.; Kuo, L.; Vacca, J. P. Structure-Based Design of Potent and Selective Cell-Permeable Inhibitors of Human SS-Secretase(BACE-). *J. Med. Chem.* **2004**, *47*, 6447−6450.

(40) Ghosh, A. K.; Kumaragurubaran, N.; Hong, L.; Kulkarni, S.; Xu, X.; Miller, H. B.; Reddy, D. S.; Weerasena, V.; Turner, R.; Chang, W.; Koelsch, G.; Tang, J. Potent Memapsin 2 (Beta-Secretase) Inhibitors. Design, Synthesis, Protein-Ligand X-Ray Structure, and In Vivo Evaluation. *Bioorg. Med. Chem. Lett.* **2008**, *18*, 1031−1036.

(41) Radi, M.; Tintori, C.; Musumeci, F.; Brullo, C.; Zamperini, C.; Dreassi, E.; Fallacara, A. L.; Vignaroli, G.; Crespan, E.; Zanoli, S.; et al. Design, Synthesis, and Biological Evaluation of Pyrazolo [3, 4-D]

DOI: 10.1021/acs.jcim.7b00128

Pyrimidines Active in Vivo on the Bcr-Abl T315I Mutant. *J. Med. Chem.* **2013**, *56*, 5382−5394.

(42) Apsel, B.; Blair, J. A.; Gonzalez, B.; Nazif, T. M.; Feldman, M. E.; Aizenstein, B.; Hoffman, R.; Williams, R. L.; Shokat, K. M.; Knight, Z. A. Targeted Polypharmacology: Discovery of Dual Inhibitors of Tyrosine and Phosphoinositide Kinases. *Nat. Chem. Biol.* **2008**, *4*, 691−699.

(43) Adams, J. A. Activation Loop Phosphorylation and Catalysis in Protein Kinases: Is There Functional Evidence for the Autoinhibitor Model? *Biochemistry* **2003**, *42*, 601−607.

(44) Cui, J. J.; Tran-Dubé, M.; Shen, H.; Nambu, M.; Kung, P. P.; Pairish, M.; Jia, L.; Meng, J.; Funk, L.; Botrous, I.; McTigue, M.; Grodsky, N.; Ryan, K.; Padrique, E.; Alton, G.; Timofeevski, S.; Yamazaki, S.; Li, Q.; Zou, H.; Christensen, J.; Mroczkowski, B.; Bender, S.; Kania, R. S.; Edwards, M. P. Structure Based Drug Design of Crizotinib (PF-02341066), a Potent and Selective Dual Inhibitor of Mesenchymal-Epithelial Transition Factor (c-MET) Kinase and Anaplastic Lymphoma Kinase (ALK). *J. Med. Chem.* **2011**, *54*, 6342−6363.

(45) Varaprasad, C. V.; Barawkar, D.; Abdellaoui, E.; Chakravarty, H.; Allan, S.; Chen, M.; Zhang, H.; Wu, J. Z.; Tam, R.; Hamatake, R.; Lang, S.; Hong, Z. Discovery of 3-Hydroxy-4-Carboxyalkylamidino-5-Arylamino-Isothiazoles as Potent MEK1 Inhibitors. *Bioorg. Med. Chem. Lett.* **2006**, *16*, 3975−3980.

(46) Wikberg, J. E. S.; Mutulis, F. Targeting Melanocortin Receptors: An Approach to Treat Weight Disorders and Sexual Dysfunction. *Nat. Rev. Drug Discovery* **2008**, *7*, 307−323.

(47) McBriar, M. D.; Guzik, H.; Shapiro, S.; Paruchova, J.; Xu, R.; Palani, A.; Clader, J. W.; Cox, K.; Greenlee, W. J.; Hawes, B. E.; Kowalski, T. J.; O'Neill, K.; Spar, B. D.; Weig, B.; Weston, D. J.; Farley, C.; Cook, J. Discovery of Orally Efficacious Melanin-Concentrating Hormone Receptor-1 Antagonists as Antiobesity Agents. Synthesis, SAR and Biological Evaluation of Bicyclo[3.1.0]hexyl Ureas. *J. Med. Chem.* **2006**, *49*, 2294−2310.

(48) Richardson, T. I.; Ornstein, P. L.; Briner, K.; Fisher, M. J.; Backer, R. T.; Biggers, C. K.; Clay, M. P.; Emmerson, P. J.; Hertel, L. W.; Hsiung, H. M.; et al. Synthesis and Structure-Activity Relationships of Novel Arylpiperazines as Potent and Selective Agonists of the Melanocortin Subtype-4 Receptor. *J. Med. Chem.* **2004**, *47*, 744−755.

(49) Xu, F.; Wu, H.; Katritch, V.; Han, G. W.; Jacobson, K. A.; Gao, Z.-G.; Cherezov, V.; Stevens, R. C. Structure of an Agonist-Bound Human A2A Adenosine Receptor. *Science* **2011**, *332*, 322−327.

(50) Clark, P.; Rowland, S. E.; Denis, D.; Mathieu, M.-C.; Stocco, R.; Poirier, H.; Burch, J.; Han, Y.; Audoly, L.; Therien, A. G.; et al. MF498 [N-{[4-(5, 9-Diethoxy-6-Oxo-6, 8-Dihydro-7H-Pyrrolo [3, 4-G] Quinolin-7-Yl)-3-Methylbenzyl] Sulfonyl}-2-(2-Methoxyphenyl) Acetamide], a Selective E Prostanoid Receptor 4 Antagonist, Relieves Joint Inflammation and Pain in Rodent Models of Rheuma. *J. Pharmacol. Exp. Ther.* **2008**, *325*, 425−434.

Figure 3.1 provided as the supporting information of the publication reports the frequent RECAP cores for the top ten PRPs of protease, kinase and GPCR. It should be noted that RECAP cores presented in **Figure 5** of the publication are a subset of the ones shown in **Figure 3.1**.

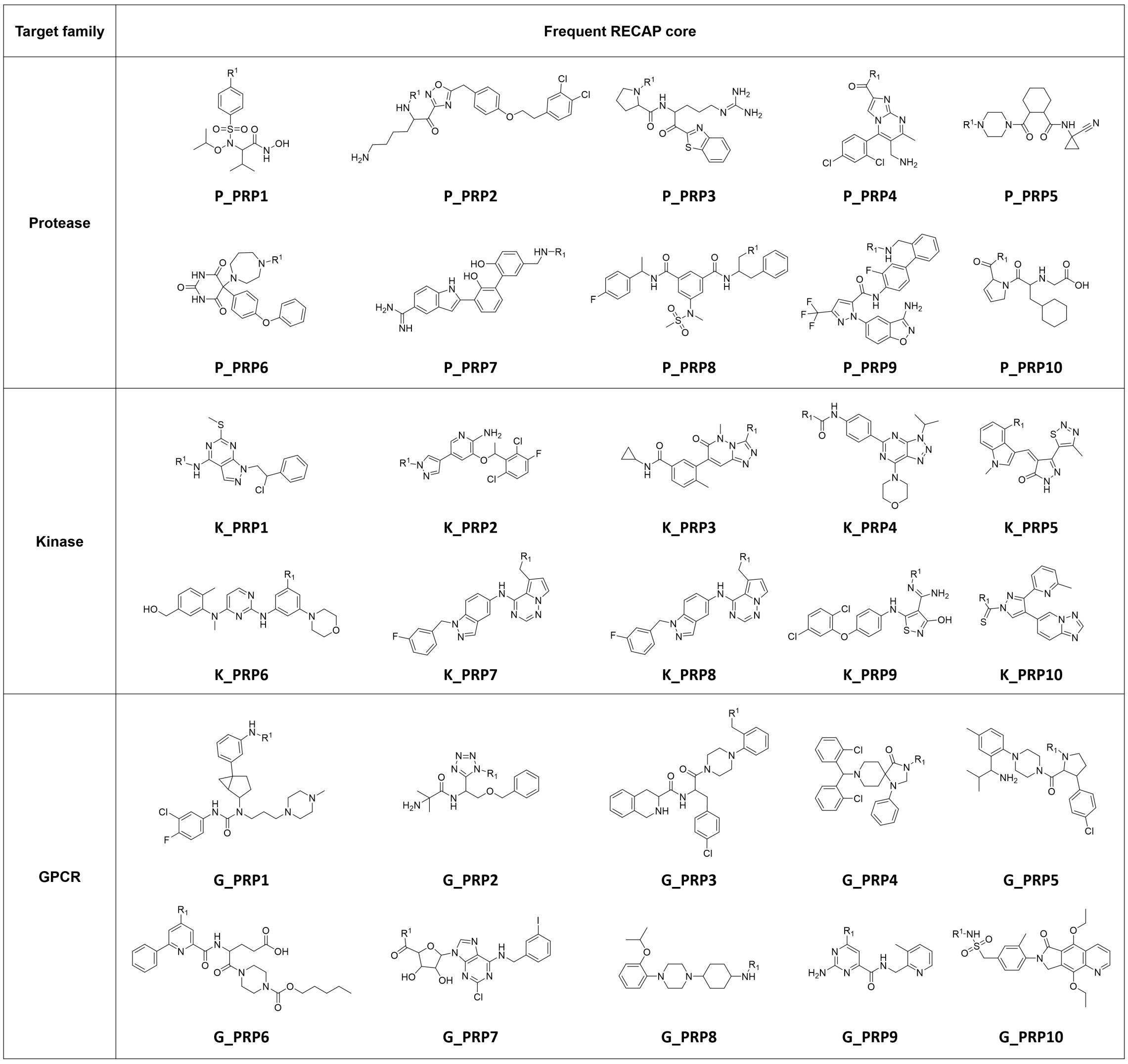


Figure 3.1: Frequent RECAP cores. The most frequent RECAP cores represented in top 10 PRPs of individual target superfamily (protease, kinase, and GPCR).

Summary

A conceptually different approach to privileged substructure extraction has been reported herein using GTM. Universal GTM was used in our analysis to map an initial pool of 152,274 compounds from ChEMBL, from which the most representative compound subsets were extracted for each of the three major target superfamilies (protease, kinase, and GPCR). GTM-based classification models were generated to assess the ability of universal GTM to discriminate ligands of different families within a single superfamily. Classification performance was consistently better than random in all cases (balanced accuracy range: 0.6 to 0.85) except for threonine protease inhibitors, which had low balanced accuracy due to its smaller set size compared to other proteases. Accurate target family-based classification suggested that the GTM-based approach was suitable to prioritize compound clusters that were representative of a target family.

A GTM landscape consists of a squared grid of “nodes” onto which compounds are projected, on the basis of fuzzy compound to node association scores (responsibilities). Each compound mapped onto a GTM is characterized by its responsibility vector and structurally related compounds tend to have similar responsibility vectors. In order to identify compound subsets representing common structural features, responsibility vectors were converted into a discretized form called responsibility patterns. Responsibility patterns were prioritized as “privileged” (PRP) towards a target superfamily if they characterized a large fraction of compounds belonging to that target superfamily but their global occurrence rate in the entire set of compounds (152,274 ChEMBL compounds) was significantly lower. Quantitatively, this was done by calculating privileged factor score for individual responsiblity pattern and assigning “privileged” status to the top ten scoring RPs for a given target superfamily. Regions of GTM pointing to PRP were delineated and corresponding compounds were extracted. Interestingly, majority of PRPs of protease and GPCR were target-specific, i.e., all compounds within a PRP belonged to single target family. However, all of the PRPs of kinase inhibitors represented at least two kinases.

PRP-specific compounds were subjected to fragmentation based on retrosynthetic rules (RECAP), giving rise to a RECAP-core and substituent for individual compound. This was followed by visual analysis of frequent RECAP-cores to identify “privileged structural motifs”. Privileged structural motifs detected from our methodology were reported in medicinal chemistry literature as characteristic features of active compounds and even approved drugs.

Our study provides a dual-purpose application of “universal GTM”:
(i) external validation of universal GTM-based models by assssing their ability to

discriminate between ligand classes for targets from different superfamilies.
(ii) analysis of GTM from medicinal chemistry perspective by assessing emerging SAR patterns. We applied GTM to delineate and prioritize regions of chemical space that were most preferred by compounds active against a given target superfamily. This paved a way to directly focus such compound subsets and subsequently extract privileged structural motifs from them. Our approach emphasizes on the idea that privileged substructures need not necessarily be defined in terms of a single molecular framework but might comprise fuzzy sets of similar scaffolds or pharmacophore-like patterns.

My major contributions to this work have been identifying PRPs for individual target superfamilies based on compound distribution and frequency analysis in universal GTMs and subsequently fragmenting the compounds within each PRP following RECAP rules. Further, I have analyzed the specificity of PRP compounds towards different target families within a given superfamily and monitored the evolution of PRPs in ChEMMBL.

Herein, we presented the application of GTM to prioritize compound clusters that are representative of a given target family and to extract of meaningful structural patterns relating to their biological activity. The probabilistic dimensionality reduction technique provided by the GTM algorithm also serves as an interesting platform for high-dimensional data visualization. In the next chapter, GTM has been utilized as a tool to visualize SAR patterns in large data sets, which is another major domain of SAR analysis.

Chapter 4

From Bird's Eye Views to Molecular Communities: Two-Layered Visualization of Structure-Activity Relationships in Large Compound Data Sets

Introduction

SAR analysis, a central task in medicinal chemistry is traditionally carried out on a case-by-case basis, i.e., focusing on one compound series at a time.[102] However, the task becomes increasingly challenging when the data set size grows beyond individual analog series.[103–106] In such cases, structure and activity need to be compared across a large and heterogeneous group of compounds, often making it nearly impossible to systematically capture all structure-activity relationships. Large-scale SAR explorations are generally accomplished using visualization tools. Activity landscape[58,106,107] represents a prime example of SAR visualization tool, and many different forms of landscape views have been proposed over the years to analyze SAR features, mainly, activity cliffs.[7,107,108] Identification of activity cliff forming compound pairs in data sets provide a valuable information for compound design and optimization. Activity cliffs are formed by structurally similar compounds with significant potency differences, a principal characteristic of SAR discontinuity.[78,87] Regions of SAR discontinuity comprise compound subsets forming activity cliffs, and thus, provide the most SAR-information.

In this study, we present a two-component visualization approach that combines a compound pair-based variant of GTM[5] with CSN[85,109] to facilitate SAR analysis in large and heterogeneous data sets. CSNs, originally introduced to represent biologically relevant subspaces of the chemical space, have been extended in various ways to visualize local SAR environments.[22,24,97,109,110] However, in case of increasingly large data sets, network representations generally become difficult to navigate. GTM, on the other hand, has proven to be a very powerful tool, with ability to map millions of compounds and visualize their activity landscapes. Global (low-resolution) activity landscape visualization provided by GTM enables identification and prioritization of SAR informative regions. The local (higher-resolution) CSN projection of prioritized regions could be used to inspect molecular communities responsible for SAR discontinuity and further resolve these regions at the level of individual compounds and their relationships.

Materials and Methods

Three K_i-based compound data sets were assembled from compound optimization efforts from BindingDB[111] database, including kappa opid receptor (KOR), adenosine a3 receptor ligands (AAR), and factor Xa inhibitors (FXA) with 1645, 1862, and 2202 compounds, respectively. These sets were comparably larger than others with high-confidence activity data and yielded 10,104, 9575, and 14,493 transformation size-restricted MMPs, respectively. If multiple K_i measurements were available for a given compound, geometric mean of all available values was taken as the final potency measurement, provided all values fell within the same order of magnitude. A large antimalarial screening set (AMS)[112] containing 13,176 compounds with confirmed inhibitory activity in malarial parasite growth assays and available XC_{50} value was selected. XC_{50} is a high-throughput screening parameter for potency, which maybe considered as an estimator of IC_{50} value for antimalarial compounds.[112] The antimalarial screening set yielded a total of 37,008 MMPs. MMPs formed between two weakly potent ($pXC_{50} < 6.5$) from screening set compounds were not of much relevance from an SAR perspective, and thus, were discarded in our analysis.

So far, GTM landscapes were generated only for individual compounds and annotated based on the their molecular properties or biological activity.[79,80,82] Compound pair-based GTMs were constructed for the first time, using MMPs as the mapping items rather than individual compounds. For each compound data set, transformation size-restricted MMPs were systematically generated using in-house implementation of Hussain and Rea[30] algorithm based on OEChem toolkit.[113] Following size restriction criteria was applied in MMP calculation to limit the chemical modification to that of an analog series: core structures were required to have at least twice the size of exchanged substituents, the size difference of exchanged fragments was limited to at most eight non-hydrogen atoms, and the maximal size of an exchanged fragment was set to 13 non-hydrogen atoms.[24]

ISIDA fragment count descriptors[37,40] were calculated for compounds participating in MMPs. MMPs were then encoded as single descriptor vector obtained by concatenating fragment counts of individual MMP partners. Concatenation involved combining the sum and difference vectors obtained by calculating the sum and absolute difference of the counts of individual fragment descriptors, respectively.[114] Concatenated descriptor vectors were used as an input for an evolutionary algorithm[81] required to optimize descriptor combinations and GTM set-up parameters to generate the GTM manifold for pair-based maps. GTM-based classification models were built on the basis of MMPs annotated with potency differences as positive and negative instances. Positive class was represented by MMP partners with at least two orders and one order of magnitude potency difference for compound optimization and screening sets, respectively. The negative instances of these compound sets included pairs with less than two orders or one order of magnitude difference in their potency, respectively. The classification models were validated using 3-fold cross-validation, which was repeated thrice after reshuffling the MMPs in the set. The model performance was evaluated by the fitness score calculated as the mean of the cross-validated balanced accuracy of each cross-validation experiment, penalized by their standard deviation.

After classification models were generated, all MMPs from a given data set was projected on the corresponding pair-based GTM for visualization. It should be noted that the classification models built in the current work was not intended for further predictive purposes. They were generated to ensure the neighborhood-behavior compliance of the selected GTMs and to use them as an SAR analysis tool.

The concept of responsibility pattern-based binning of items mapped onto a GTM landscape is applied herein to group MMPs on the selected pair-based GTMs. For each responsibility pattern, an SAR discontinuity score was calculated as follows:

$$SARDisc(RP) = \frac{\sum\limits_{i}^{m} \sum\limits_{i,j>i}^{m} |pot_{\mathrm{i}} - pot_{\mathrm{j}}|}{N} \quad \forall\, i, j \rightarrow MMP \tag{4.1}$$

where i and j are compounds forming an MMP, pot_{i} and pot_{j} are their respective potencies, m is the total number of compounds, and N the total number of MMPs for the RP. $SARDisc(RP)$ is reminiscent of the discontinuity score component of the SAR index[87] and accounts for all pairwise potency differences between compound pairs forming an MMP associated with a given RP. The score is high for RPs associated with structural analogs having large potency variations. Pair-based GTMs were colored on the basis of RP-specific SAR discontinuity scores.

Compounds inhabiting individual RP were extracted to generate MMP-CSNs in which pairwise substructure relationships between MMPs could be intuitively followed through network representation. Nodes of MMP-CSNs corresponded to compounds and were connected by edges if they formed pairwise MMP relationships.[97] In addition, nodes were colored by potency of the compound they represent using continuous color spectrum from red (lowest potency in the data set) over yellow (intermediate) to green (highest potency). MMP-CSNs were built using Java software and the Java universal network/graph framework (JUNG)[115] and their layout was generated using the Fruchterman-Reingold algorithm,[116] which organizes similar objects into clusters and separates clusters for display in a force-directed manner. MMP relationships that were associated with a given RP were highlighted in corresponding RP-specific MMP-CSN to provide local SAR views.

Results and Discussion

Two-layered visualization approach, combining GTM and CSN representations, was introduced herein to facilitate SAR exploration in increasingly large compound data sets. The first layer was provided by pair-based GTMs, representing global or bird's eyes views of activity landscapes. Regions of the map pointing to high SAR discontinuity, i.e., SAR-informative zones were identified and compounds populating them were extracted. Prioritized compounds were analyzed in detail at the level of individual compounds in MMP-CSNs.

Figure 4.1 shows best performing pair-based GTMs for three compound optimization sets and one antimalarial screening set, colored by SAR discontinuity scores. The GTM set-up parameters and descriptors used to build these maps are detailed in **Table 4.1**.

Data set	**Descriptors**	**Size**	**Num. RBF**	**RBF width**	**BA**
KOR	IIRA–P-1-6: ISIDA counts of atom centered fragments based on sequences of atoms of fixed length, covering the first and sixth coordination sphere	24	12	1.1	0.76
FXA	IAB-PH-P-2-6: ISIDA pharmacophore annotated counts of atom pairs found two to six bonds apart, including information about bonds nearest to terminal atoms	26	11	1.4	0.71
AAR	IIRA–P-1-6: ISIDA counts of atom centered fragments based on sequences of atoms of fixed length, covering the first and sixth coordination sphere	29	11	1.1	0.71
AMS	IIA–P-1-3: ISIDA counts of atom centered fragments based on sequences of atoms, covering the first and third coordination sphere	27	13	2.5	0.73

Table 4.1: Descriptors and parameters of top-scoring GTMs. For the best GTMs obtained for each data set, selected descriptors and balanced accuracy (BA) resulting from an evolutionary optimizer are reported. In addition, "Size" refers to the number of nodes defining the edge length of the square grid representing the GTM. Furthermore, the number of radial basis functions (Num. RBF) and their width are reported.

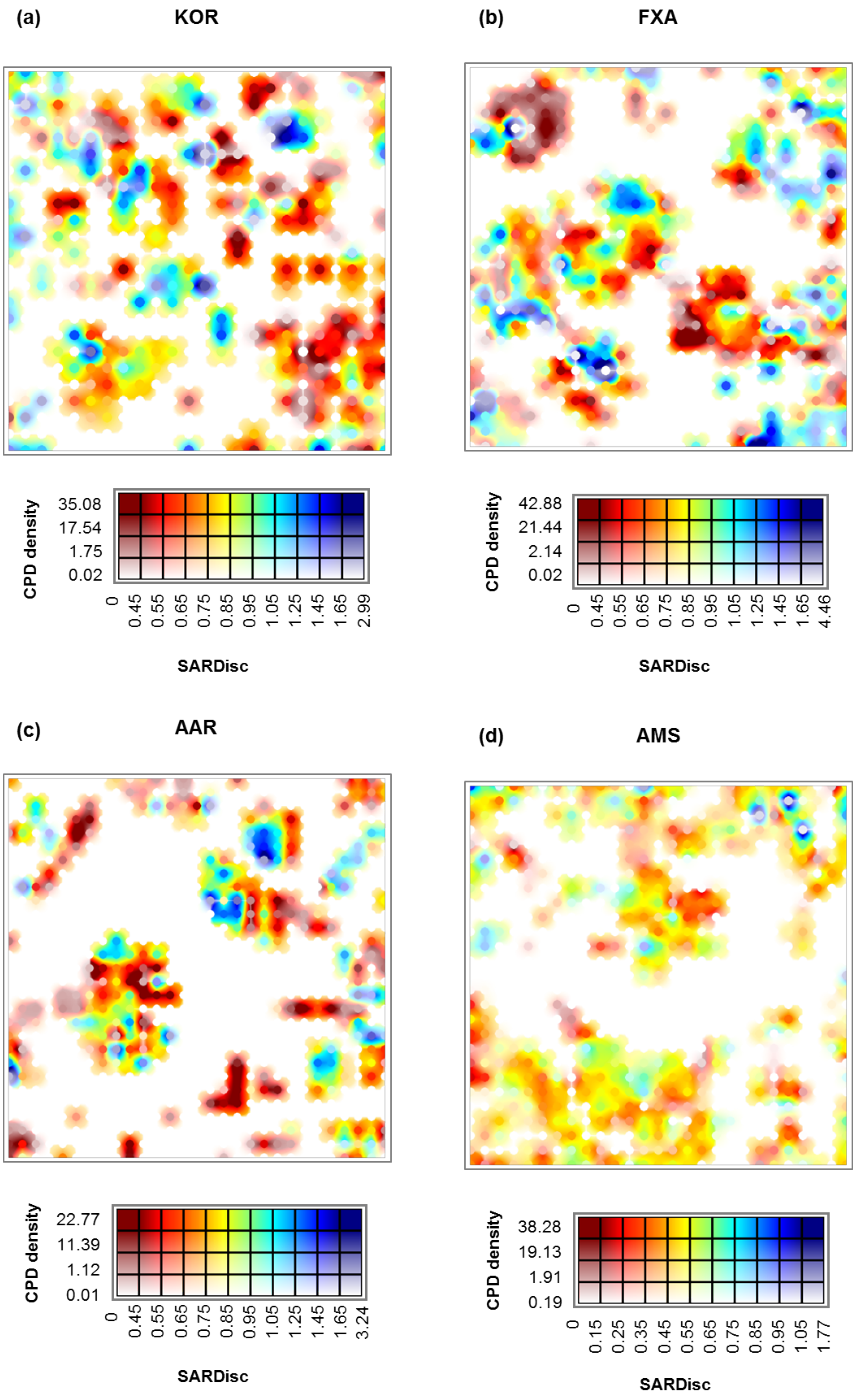


Figure 4.1: GTM projections. For all compound data sets, GTM 2D maps with RPs colored by SAR discontinuity score are shown. The maps were based on classification models. As indicated in the color spectrum, increasing transparency indicates decreasing compound density of RPs. Maps were colored using a continuous spectrum from red (lowest discontinuity score) through yellow and green (intermediate) to blue (highest score). (a) KOR, (b) FXA, (c) AAR, (d) AMS.

RPs of KOR were evenly distributed across the map unlike those of FXA and AAR, which were notably clustered and separated from each other. Maps of AMS sets were characterized by a low degree of SAR discontinuity, which could be attributed to presence of many weakly potent hits with different structures. Nevertheless, nodes representing high SAR discontinuity were observed in the upper right corner of the map. In each of the maps, different SAR environments were clearly distinguishable and several zones of high SAR discontinuity could be spotted. This provided enough evidence of the ability of pair-based GTMs to differentiate between regions of varying local environments.

In order to prioritize responsibility patterns, their SAR discontinuity scores were monitored for all compound data sets, as shown in **Figure 4.2**. Only RPs associated with at least 10 compounds were taken into account as smaller subsets of compounds are not very SAR informative. In each set, RPs with discontinuity score of at least one standard deviation above the mean of the distribution were prioritized for CSN analysis.

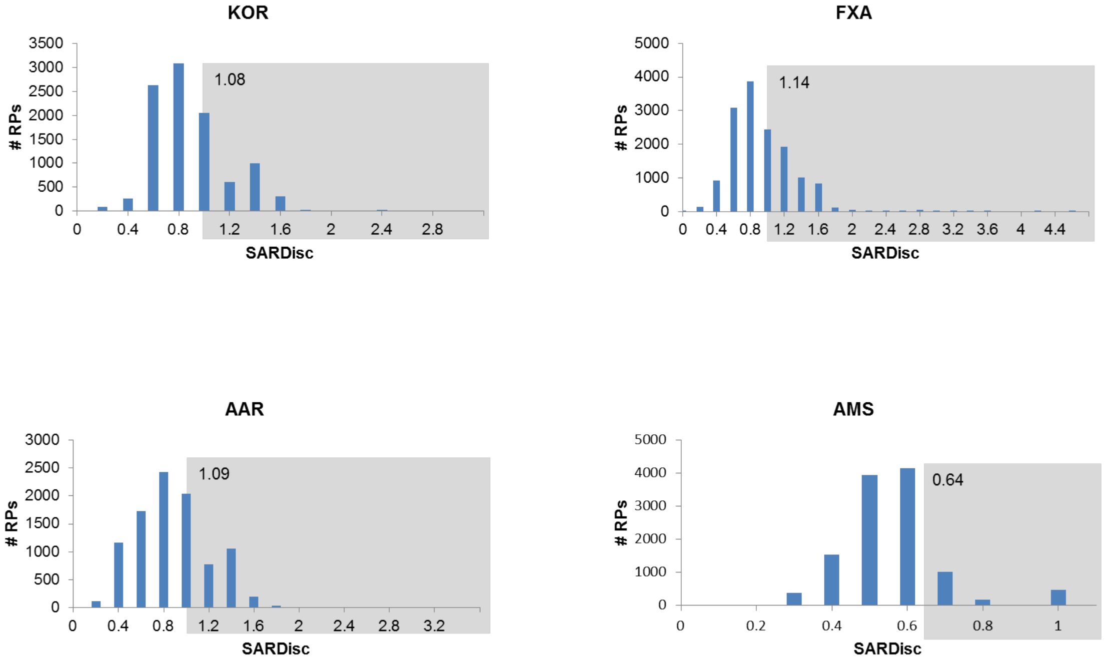


Figure 4.2: SAR discontinuity of RPs. Shown is the distribution of RPs over SAR discontinuity scores (SARDisc) for the four compound sets. RPs reaching or exceeding a SARDisc value of one sigma above the mean of each distribution are shown on a gray background.

Compounds representing prioritized RPs were used to generate RP-specific

MMP-CSNs, thus shifting the focus of our analysis from pair-based display to individual compounds and their pairwise structural relationships. Further, the change from GTM to CSN allowed transition from global views that helped to pre-select compound subsets of interest to more localized SAR views in which these compounds could be scrutinized in more detail. Since MMP-CSNs were generated exclusively for compound subsets specific to a given RP, they were easy to navigate and interpret.

Figure 4.3 compares CSNs of compound subsets from top-scoring RPs of each data set and illustrates how local SAR environments were distinguished. For KOR, CSN 1 and 2 contained densely connected central clusters of differently colored nodes corresponding to structurally very similar compounds with large potency variations (**Figure 4.3a**), indicating highly discontinuous and SAR-informative subsets of active compounds. By contrast, CSN 3 was characterized by the presence of many compounds with comparable intermediate potency and one weakly potent compound that was similar to many others (**Figure 4.3a**). This constellation also gave rise to a high discontinuity score, due to many pairwise comparisons involving the weakly potent analog, but was much less SAR-informative than the compound clusters in CSN 1 and 2. The CSNs for FXA in **Figure 4.3b** also revealed different SAR characteristics. CSN 1 contained many pairs of compounds with moderate potency variations, which did not convey useful SAR information and one cluster with compounds having large potency variations. However, this cluster was only sparsely connected, indicating the presence of only partly similar compounds, which is also limited SAR information. More informative clusters were observed in CSN 2, which contained only relatively few compounds, and especially in CSN 3. This CSN contained weakly similar compounds, but also a densely connected cluster with SAR-informative compounds. Comparable observations were made for AAR CSNs shown in **Figure 4.3c**. Here, prioritized local SAR environments also differed in their composition and characteristics, with decreasing SAR information content from CSN 1 (two densely connected clusters of comparable size) over CSN 2 (one densely connected cluster) to CSN 3 (none). Moreover, despite the prevalence of typical screening hits, AMS also contained small subsets of compounds with high SAR information content, as revealed by the CSNs in **Figure 4.3d**. The identification of these SAR islands within the screening data background and their characterization, as further discussed below, was straightforward using GTM-CSN analysis, which we considered an encouraging finding.

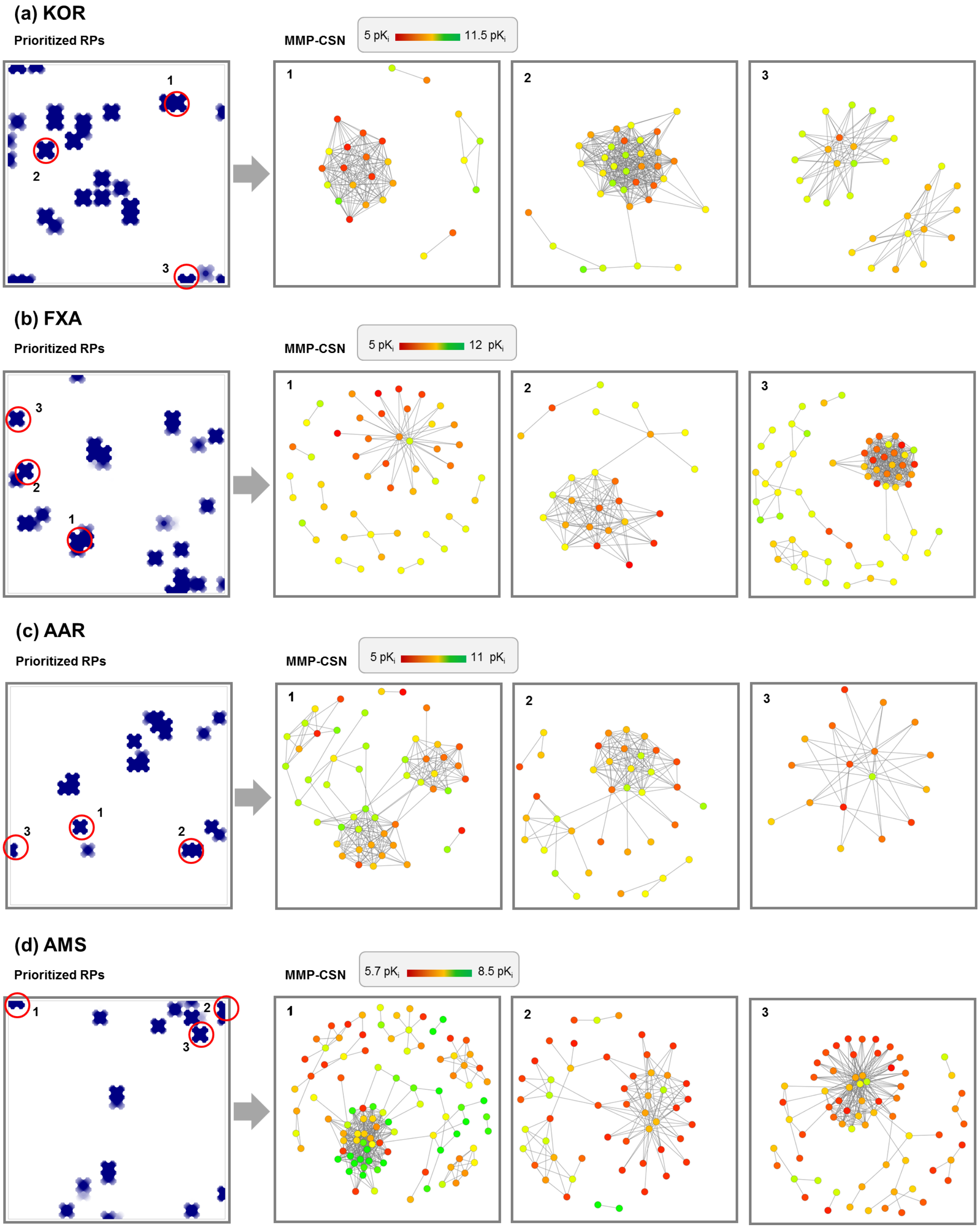
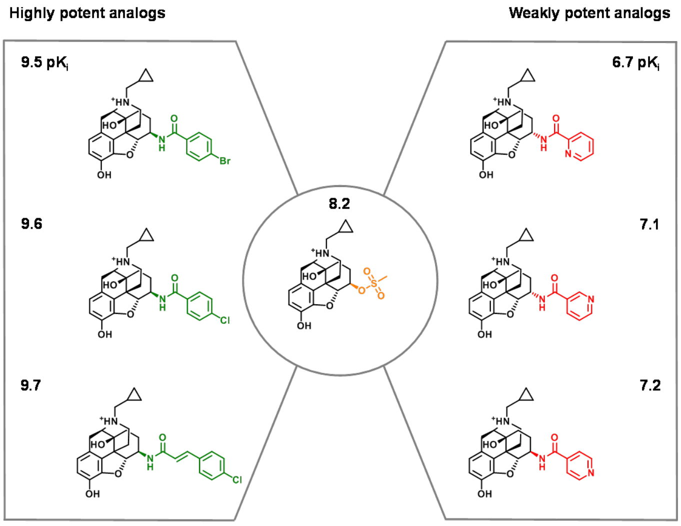
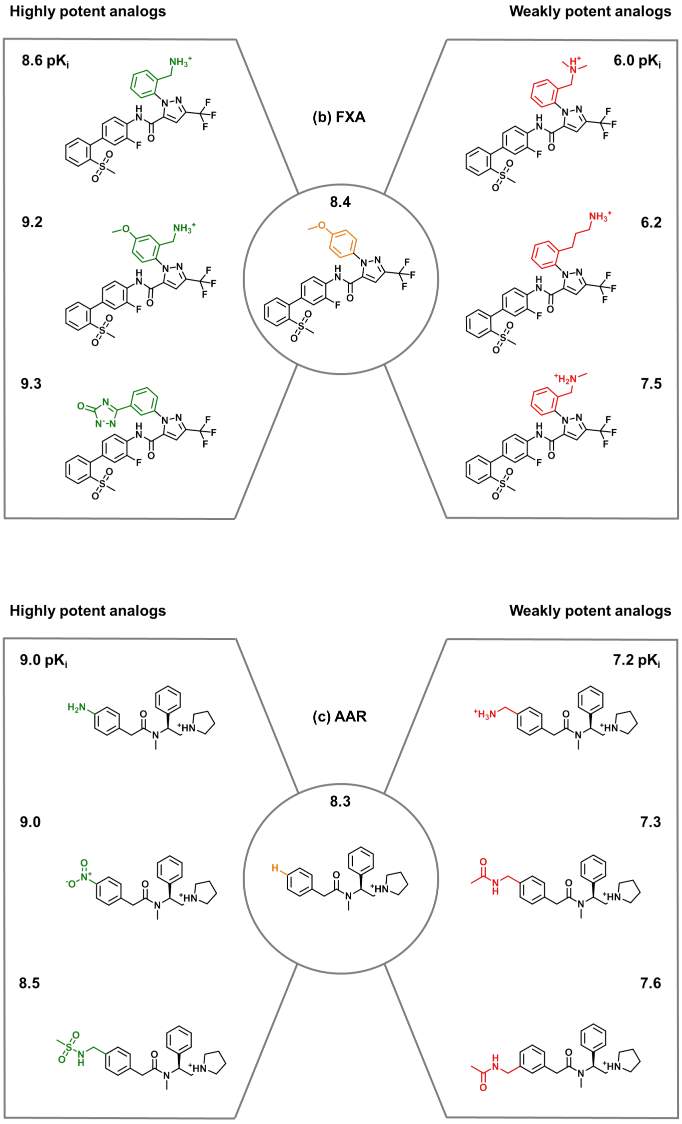


Figure 4.3: GTM-CSN views. RPs shown on a simplified GTM map were prioritized on the basis of SAR discontinuity. For corresponding compound subsets, CSN representations with constant layout were generated in which nodes were colored by compound potency. Comparison of these CSNs made it possible to distinguish between different local SAR environments. The orders of CSNs indicate decreasing SAR discontinuity scores of corresponding RP. (a) KOR, (b) FXA, (c) AAR, (d) AMS

MMP-CSNs of high-priority RPs were interactively navigated to search for analog series displaying interesting SAR patterns. For each of these series of analogs, structural modifications leading to potency variations can be explored as shown in **Figure 4.4**. Representative structures from KOR set in **Figure 4.4a** focus on a substitution site in a series of analogs where the introduction of bromo- and chloro- benzamides increased potency compared to the reference compound, whereas pyridinecarboxamides reduced potency by at least one order of magnitude. Similar observations were made for different R-groups in FXA (**Figure 4.4b**) and AAR structures (**Figure 4.4c**). Examples of AMS set in **Figure 4.4d** shows that the presence of phenyl, substituted phenyl, and bulky alkyl groups at the designated substitution site led to increase in potency, whereas the introduction of small substituents such as methyl or ether groups resulted in weakly potent analogs.




Highly potent analogs
Weakly potent analogs
8.6 pK$_i$
9.2
9.3
(b) FXA
8.4
6.0 pK$_i$
6.2
7.5
Highly potent analogs
Weakly potent analogs
9.0 pK$_i$
9.0
8.5
(c) AAR
8.3
7.2 pK$_i$
7.3
7.6

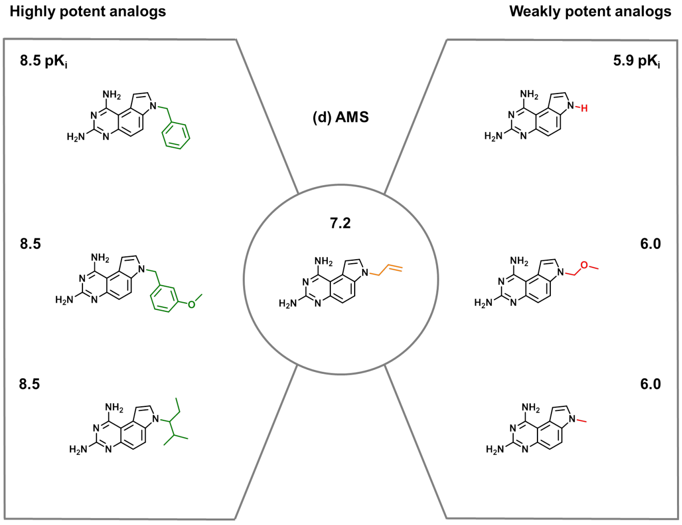


Figure 4.4: SAR-relevant chemical changes. For exemplary analogs from CSNs of prioritized RPs, chemical modifications (transformations) are shown that led to increases or decreases in potency.

Conclusions

In this study, a two-component visualization tool to explore SAR patterns in large and heterogeneous data sets has been introduced. The challenge associated with SAR analysis in data sets containing multiple series of compounds has be addressed by first combining a global and local views of the activity landscape. The global view guided the analysis by pointing towards areas of interest, in our case, SAR discontinuous zones. Once desired SAR-informative regions were delineated, they were analyzed in greater detail in high-resolution local views provided by CSN representations. Since CSN were focused towards prioritized compound subsets characterized by high discontinuity scores, it was much easier to navigate and identify interesting SAR patterns. In our proof-of-concept study, the GTM-CSN approach was applied to compound optimization and screening data sets. GTMs capable of differentiating between different local SAR environments, from which compound subsets inhabiting SAR-informative regions were captured and projected onto MMP-CSNs. CSNs further helped to resolve these regions at the level of individual compounds and their relationships. Above all, our findings suggest that the combined visualization approach proposed by GTM-CSN provides a meaningful extension to currently available SAR visualization techniques.

My major contributions to this work have been identifying MMPs formed by ligands active against individual targets and encoding them as single descriptor vector based on ISIDA fragment counts of each MMP partner. My contributions also include generation of compound pair-based GTM classification landscapes, extraction of RPs pointing to SAR discontinuous regions, and generation of RP-specific MMP-CSNs.

Activity cliffs represent the most prominent features of activity landscapes and is an attractive concept for large-scale SAR analysis. Activity cliffs have thus been extensively studied by applying data mining as well as visualization methods. In the next chapter, computational models for predicting activity cliffs have been discussed.

Chapter 5

Prediction of Activity Cliffs Using Condensed Graphs of Reaction Representations, Descriptor Recombination, Support Vector Machine Classification, and Support Vector Regression

Introduction

Activity cliffs are defined by pairs of structurally similar compounds with significant potency difference. As such, they represent extreme discontinuity in SAR landscape. They are of paramount importance in SAR exploration as rich source of SAR information since they reveal small chemical modifications that lead to significant change in biological activity of compounds.

Two fundamental aspects need to be considered while defining an activity cliff, i.e., the structural similarity and potency difference assessment criteria. Numerical measures of structural similarity, such as, Tanimoto coefficient, calculated using molecular fingerprints is widely used for determining structural similarity between activity cliff partners. However, calculated similarity heavily relies on the type of molecular fingerprints as well as the similarity metric applied. Furthermore, numerical similarity values are not always easily interpretable from medicinal chemistry point of view. In contrast, MMPs provide a chemically intuitive means to assess structural similarity between compounds. Since, compounds forming an MMP differ by chemical change only at a single site, relating structural changes to potency is relatively easier. More recently, transformation size-restricted MMPs have been applied as structural similarity criterion in activity cliff formation, giving rise to so-called MMP-cliffs. We have applied MMP-based definition of activity cliff in this study.

Activity cliffs have been extensively explored and many studies have been reported so far, mainly focusing on mining of activity cliffs in compound data set and their graphical representations. Predictions of small structural modifications that result in large shifts of biological activity could yield significant inferences for compound design and optimization. However, only a few studies have attempted to predict if a given pair of structurally similar compounds form an activity cliff or not. Attempts made so far in this direction have employed machine learning methods such as, support vector machines or random forests. In this study, we have followed two different approaches to represent MMPs - condensed graph of reaction, originally developed for modeling chemical reactions and plain descriptor recombination, a strategy used for QSPR modeling of nonadditive mixtures. Single descriptor vectors encoding individual MMPs were derived from these representations and given as an input for support vector machine models. Support vector classification was used for predicting activity cliff/non-cliff status, whereas support vector regression models predicted the direction of the potency increase or decrease of a given compound pair.

JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

pubs.acs.org/jcim

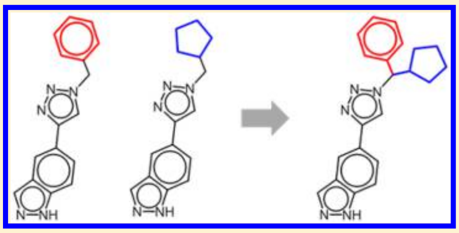
Prediction of Activity Cliffs Using Condensed Graphs of Reaction Representations, Descriptor Recombination, Support Vector Machine Classification, and Support Vector Regression

Dragos Horvath,[†] Gilles Marcou,[†] Alexandre Varnek,*[,†] Shilva Kayastha,[†,‡] Antonio de la Vega de León,[‡] and Jürgen Bajorath*[,‡]

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

[‡]Department of Life Science Informatics, B-IT, LIMES Program Unit Chemical Biology and Medicinal Chemistry, Rheinische Friedrich-Wilhelms-Universität, Dahlmannstrasse 2, D-53113 Bonn, Germany

S Supporting Information



ABSTRACT: Activity cliffs (ACs) are formed by structurally similar compounds with large differences in activity. Accordingly, ACs are of high interest for the exploration of structure−activity relationships (SARs). ACs reveal small chemical modifications that result in profound biological effects. The ability to foresee such small chemical changes with significant biological consequences would represent a major advance for drug design. Nevertheless, only few attempts have been made so far to predict whether a pair of analogues is likely to represent an AC—and even fewer went further to quantitatively predict how "deep" a cliff might be. This might be due to the fact that such predictions must focus on compound pairs. Matched molecular pairs (MMPs), defined as pairs of structural analogs that are only distinguished by a chemical modification at a single site, are a preferred representation of ACs. Herein, we report new strategies for AC prediction that are based upon two different approaches: (i) condensed graphs of reactions, which were originally introduced for modeling of chemical reactions and were here adapted to encode MMPs, and, (ii) plain descriptor recombination—a strategy used for quantitative structure−property relationship (QSPR) modeling of nonadditive mixtures (MQSPR). By applying these concepts, ACs were encoded as single descriptor vectors used as input for support vector machine (SVM) classification and support vector regression (SVR), yielding accurate predictions of AC status (i.e., cliff vs noncliff) and potency differences, respectively. The latter were predicted in a compound order-sensitive manner returning the signed value of expected potency differences between AC compounds.

INTRODUCTION

Activity cliffs (ACs) are formed by pairs or groups of active compounds that are structurally similar or analogous and have large differences in potency.[1,2] Accordingly, the application of specific molecular similarity and potency difference criteria is of critical relevance for a consistent assessment of ACs.[1,2] Tanimoto similarity using various molecular fingerprints[3,4] has typically been calculated to define ACs structurally.[1] More recently, the matched molecular pair (MMP) formalism[5,6] has also been applied as a similarity criterion for AC formation. An MMP is defined as a pair of compounds that are only distinguished by a chemical change at a single site,[5] i.e. the exchange of a substructure, often termed a chemical transformation.[7] On the basis of MMPs with predefined transformation size restrictions, which generally limit MMPs to pairs of structural analogs,[8] ACs have been defined as pairs of compounds that form a transformation size-restricted MMP and have a potency difference of at least 2 orders of magnitude, leading to the introduction of MMP-cliffs.[8] This chemically intuitive MMP-based definition of ACs is particularly relevant for AC analysis in medicinal chemistry where ACs are of prime interest as sources of structure−activity relationship (SAR) information,[2] given their "small chemical change—large biological effect" phenotype. In addition, in drug design and chemical informatics, ACs are also of considerable interest as test cases for potency prediction and SAR analysis methods, especially given the intrinsic similarity of AC compounds. Moreover, the ability to predict ACs—or the directionality of potency changes for AC partners—would have high potential for practical applications aiming at the identification of novel potent compounds. Different from conventional compound activity predictions, focusing on ACs presents us with the challenge to predict compound pairs, rather than individual test compounds. This requires the design and implementation of new prediction schemes for machine learning, which may explain why only a few attempts have been made thus far to predict ACs. Initially, random forests,[9] i.e. arrays of decision trees, were applied to predict ACs in compound data sets.[10] Specifically, random forest models were trained to predict if a test compound would form ACs with other data set compounds, yielding moderately accurate predictions.[10] Fur-

Received: June 16, 2016
Published: August 26, 2016

DOI: 10.1021/acs.jcim.6b00359

thermore, particle swarm optimization,[11] a population-based global search technique mimicking coordinated (social) swarm formation and behavior, was applied to search data sets for ACs formed within groups of structurally similar compounds.[12] Moreover, support vector machine (SVM) classification[13,14] was used to predict ACs (MMP-cliffs) and distinguish them from non-ACs (MMPs encoding small potency differences).[15] These SVM predictions were enabled through the introduction of specifically designed kernel functions that utilized MMP core and transformation information and yielded overall accurate predictions.[15] Finally, in a conceptually related study, it was attempted to predict the direction of MMP-associated potency changes[16] using SVM regression,[17] an extension of SVM classification to predict numerical activity values. Apart from these studies, no other AC predictions have so far been reported.

In order to model properties related to MMPs, different strategies can be considered. For example, special "MMP kernels" might be designed, as reported previously,[15,16] to facilitate predictions using kernel methods; alternatively, more simplistic procedures to encode MMPs as various molecular descriptor vectors might be envisioned, which would make it possible to apply any other machine learning methods. The latter strategy represents the focal point of the current work.

Herein, we present AC prediction models based upon two different MMP representations using (i) descriptors derived from condensed graphs of reaction (CGR), which were originally introduced for modeling of chemical reactions,[18,19] and (ii) plain descriptor recombination—a strategy used for QSPR modeling of nonadditive mixtures.[20] Both of these strategies were applied to ultimately encode MMPs as vectors of a variety of fragment descriptors that are applicable to regular molecular graphs as well as CGRs. The resulting descriptor vectors provided the basis for SVM classification and SVR following a variety of advanced modeling strategies. The methodological framework for these AC predictions is detailed in the following and the results of test calculations are reported.

CONCEPTS, METHODS, AND MATERIALS

Activity Cliffs. For our analysis, ACs were defined as MMP-cliffs,[8] i.e. compound forming a transformation size-restricted MMP having a potency difference of at least 2 orders of magnitude (100-fold). Compound pairs with potency differences between one and 2 orders of magnitude are generally not considered ACs.[2] Hence, they should best be excluded from classification of ACs vs non-ACs. Furthermore, in classification, boundary effects can be controlled by excluding such compound pairs and defining non-ACs to be represented by MMPs with compounds having a potency difference of less than 1 order of magnitude (10-fold). This strategy was applied herein for AC vs non-AC classification where these class-labeled instances should be considered as discrete potency difference states rather than a continuum. However, for regression modeling, as detailed below, compound pairs across all potency difference ranges including one to 2 orders of magnitude were used (representing a departure from conventional AC definitions). MMPs were generated using in-house implementation of the algorithm by Hussain and Rea[7] by limiting fragmentation to single exocyclic bonds and applying the following transformation size restrictions.[8] Specifically, the MMP core was required to be of at least twice the size of exchanged substructures that were not permitted to be larger than 13 heavy atoms and differ by more than eight heavy atoms. Applying these size restrictions generally produces MMPs that represent typical analogs.[8]

Compound Data Sets. The compound data sets were the same as used in the previous SVM-based prediction of ACs.[15] This made it possible to directly compare the results for the AC classification reported herein with the previous SVM-based classification, as detailed below. Compounds active against seven different targets were assembled from BindingDB.[21] These seven selected K_i value-based target sets were selected to yield large numbers of MMP-cliffs formed by compounds with at least 10 μM potency. If more than one K_i measurement were available for a compound, the geometric mean of available values was calculated as the final potency annotation, provided all the values fell within the same order of magnitude; otherwise, the compound was discarded. Table 1 reports the composition of the data sets used herein, which included five different sets of enzyme inhibitors and two sets of receptor ligands containing between 58 and 2202 compounds.

Table 1. Compound Data Sets, MMP, and AC Statistics[a]

data set	CPDS	MMPs	ACs	non-AC MMPs
Factor Xa inhibitors (fxa)	2202	14493	1161	10108
Thrombin inhibitors (thr)	2037	9585	1103	6390
Cathepsin b inhibitors (catb)	150	681	120	451
Kappa opioid receptor ligands (kor)	1645	10104	649	7190
Adenosine A3 receptor ligands (aa3)	1862	9575	681	6752
Calpain 2 inhibitors (cal2)	121	1206	387	718
Janus kinase inhibitors (jak2)	58	366	109	186
total	8075	46010	4210	31795

[a] The composition of the data sets, resulting numbers of MMPs, activity cliffs, and qualifying noncliff MMPs are reported.

Modeling Strategies. Categorical and quantitative (regression) modeling were carried out. The categorical approach calibrated binary predictors able to return 1 if an MMP was predicted to be an AC and 0 otherwise. This categorical approach was considered "order-independent (OI)" because the result did not depend on the *order* in which the two members of an MMP were presented. By contrast, quantitative regression is more challenging as it requires predicting by how much the potency of the first compound of the MMP differs from the one of the second. Therefore, these predictions are "order-dependent (OD)". The methodological summaries provided in the following are complemented by detailed descriptions provided as Supporting Information.

Representing MMPs as Condensed Graphs of Reaction. CGRs represent an effective way to synthetically characterize a chemical transformation using a single molecular graph including both conventional bonds (e.g., single, double, aromatic, etc.) and "dynamical" bonds characterizing chemical transformations.[18] Formal "pseudo bond orders" are assigned to reflect the nature of the change, i.e., breaking or forming a bond or changing the bond order. Therefore, a CGR captures within a single graph all information required to specify a given reaction or, more generally, chemical transformation. Therefore, MMP formation can also be effectively encoded using CGRs, as illustrated in Figure 1a. The CGR contains two "dynamical bonds" that capture the transformation leading to cliff formation. Typically used SMILES or SMARTS[22] representations of MMPs were converted into CGRs using a

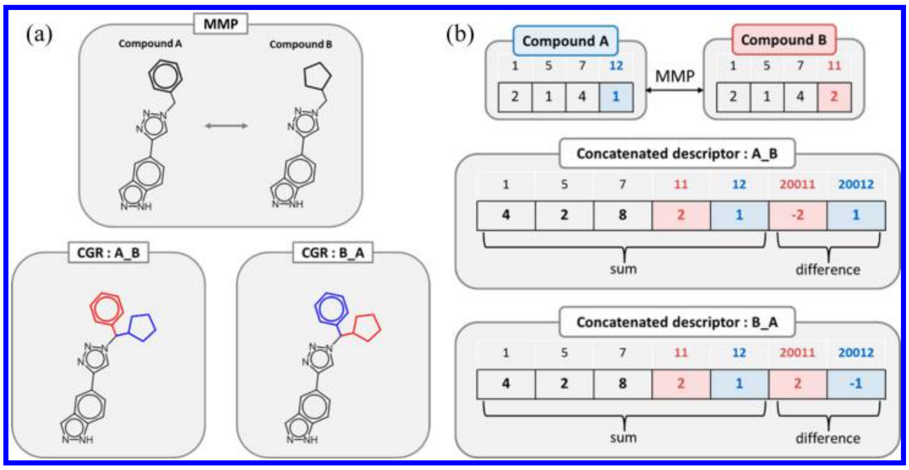


Figure 1. Alternative MMP representations. (a) CGR representation of an MMP. Broken and formed dynamical bonds and corresponding transformations are colored red and blue, respectively. Considering the order-dependence of dynamical bond deletion and formation results in two alternative CGRs. (b) Concatenated MMP descriptor vector for a pair of compounds A and B. The first vector elements, numbered from 1, represent the sum of compound vectors, while the differences are stored, in elements with indices beyond 20 000 (an arbitrary offset chosen to be larger than the dimension of the compound vector space). The first populated difference element is 20 011, because previous fragments 1, 4, and 7 are equally numerous in both compounds.

dedicated in-house java tool based upon the ChemAxon Application Programming Interface.[23] The java tool also regenerates the structures of the compounds forming the cliff and retains core and substituent information. For descriptor calculation (vide infra), structures were standardized using a web server for virtual screening[24] (see the Supporting Information for further details).

Descriptor Calculation. Descriptor vectors encoding MMPs captured in CGRs must be compliant with the order independency or dependency. Therefore, the classification and regression models require separate descriptor sets. ISIDA Fragmentor[25,26] was used to generate a large pool of alternative fragment descriptors for modeling via context-dependent fragmentation of input structures. These descriptors included sequences of circular fragments with or without specifying chemical bonds, defining dynamical bonds, or marking atoms with special status.[25] Well performing descriptor sets for classification and regression were selected using evolutionary model optimization procedures (vide infra).

Descriptors for OI-classification models were extracted from order-independent condensed graphs (OICG). These contained two dynamical bonds of same type, with marked head atoms.

For OD-regression, two distinct strategies were used to generate descriptors for OD-regression approaches:

(i) Descriptors were derived from specific OD condensed graphs (ODCG). In this case, explicit inclusion of bond information was mandatory because the order of substituents may only be inferred from dynamical bond orders.

(ii) Descriptors were also generated following QSPR strategies[20] for nonadditive mixtures. These are derived from regular ISIDA fragment descriptor sets for two individual MMP compounds. Figure 1b illustrates concatenation of descriptors for MMPs. The descriptor vector is depicted in a fingerprintlike representation in which each element counts the occurrence of a specific fragment. The first part of the concatenated descriptor is given by the sum of the corresponding fragment counts for MMP compounds and, hence, represents the entire MMP including its core and transformation. In addition, the second part reports the difference between the fragment counts and thus accounts for the transformation. It should be noted that resulting descriptors are order-dependent (the difference part will change signs when the order of the MMP compounds is reversed).

Model Building. For each ISIDA fragmentation scheme, property-descriptor matrices were generated for compound data sets. The property value in the first column of these matrices was either the potency difference Δ per se or the categorical "AC status" (i.e., cliff vs no cliff). In the OD case for regression, property—descriptor matrices systematically enumerated MMPs with alternative ordering (see the Supplementary Methods for further details).

The evolutionary optimizer[27] of the *libsvm* SVM library[28] was used to search for optimally cross-validated predictive models for each modeling strategy applied to each of the seven compound data sets. The method simultaneously optimizes descriptor set selection (out of provided candidate descriptor files) and operational parameters of SVM models, including kernel choice (out of standard *libsvm* options: linear, third-order polynomial, radial basis function, and sigmoid kernels). All OICG-derived fragmentation schemes were permitted to compete in a Darwinian quest for top validating classification models. Regression models were challenged in two distinct ways based only on MQSPR descriptors or ODCG-based fragment counts as competitors (see the Supporting Information).

The fitness function used to prioritize modeling protocols described above intrinsically relies on conventional p-fold "item-out" cross-validation. At every instance, $1/p$ of the entire MMP set is kept aside for validation of a model trained on the remaining $(p - 1)/p$ parts of the data. At $p = 3$, the cross-validation exercise used to evaluate model fitness is also repeated several (here, five) times, after reshuffling, in order to render its outcome independent of the actual regrouping of MMPs for training and test. However, MMPs left out for

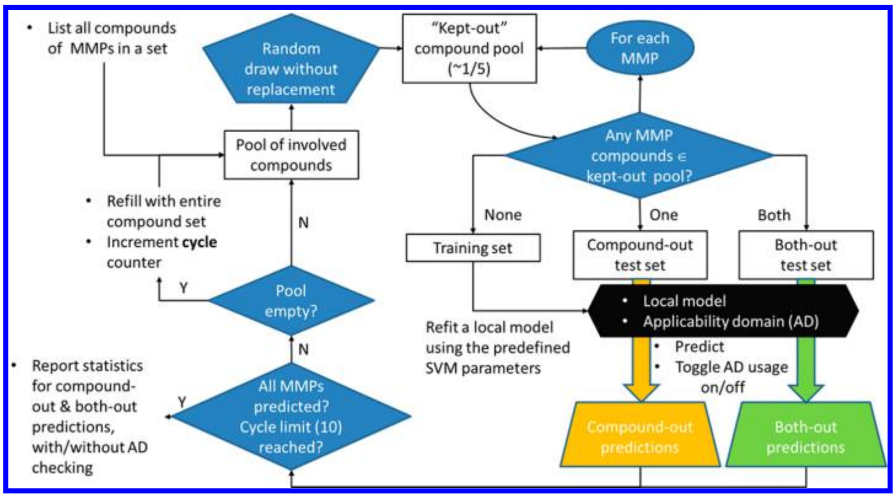

• List all compounds of MMPs in a set
Random draw without replacement
"Kept-out" compound pool (~1/5)
For each MMP
Pool of involved compounds
Any MMP compounds ∈ kept-out pool?
• Refill with entire compound set
• Increment cycle counter
N
None
One
Both
Training set
Compound-out test set
Both-out test set
Y
Pool empty?
• Local model
• Applicability domain (AD)
Refit a local model using the predefined SVM parameters
N
• Predict
• Toggle AD usage on/off
• Report statistics for compound-out & both-out predictions, with/without AD checking
Y
All MMPs predicted? Cycle limit (10) reached?
Compound-out predictions
Both-out predictions

Figure 2. Advanced cross-validation. The workflow illustrates the AXV scheme, based on (maximally) 10 cycles of 5-fold stochastic leaving out of individual compounds. The current "kept-out" compound list unambiguously defines the status of each MMP, which may be either a training set member (none of its compounds among the left-out) or assigned to one of the test sets, in order to be predicted by a local model guaranteed not to have encountered (one or both) of its compounds at training stage.

prediction may nevertheless consist of two compounds that were present in some other MMPs of the training set; although not together, or—for OD regression models—not in the same order.

Thus, while the "item-out" cross-validation results reported in this paper will be the ones obtained at model selection stage, a further, "advanced" scheme was considered, including "compound-out"[29] and "both-out" validation strategies. Figure 2 shows a workflow summarizing this advanced cross-validation (AXV) procedure, which indirectly manages the MMP set partition by directly controlling the status of individual compounds. AXV was applied to top selected models, using descriptor sets and *libsvm* operational parameter setups resulting from the evolutionary procedure. At any time, roughly one-fifth of the involved compounds are (stochastically) marked as kept-out, which triggers the implicit exclusion of all MMPs containing at least one of these compounds from the local training set. Such MMPs will be redirected to "compound-out" or "both-out" test sets, depending on whether only one or both of their constituents are tagged as kept-out. Note that keeping one-fifth of the compounds out typically results in two-thirds of all MMPs being assigned to training, while the remaining one-third is split between "compound-out" and "both-out" test sets.

To further analyze the predictive ability of the models, boundaries were defined in descriptor space as a hyper-parallelepiped to delineate applicability domains (AD). Boundaries were set by using the minimum and maximum descriptor values of the training sets instances. Since the descriptor space was defined by the structural fragments of compounds, ADs corresponded to so-called fragment control applicability domain.[30]

Control calculations for regression were carried out using random forest (RF) modeling[9] using a protocol following an approach for prediction of MMP-encoded property changes.[31]

For RF calculations, the R[32] package randomForest[33] was used. For training, an MMP was represented as the difference in values of a set of 51 2D numerical descriptors[31] calculated with the Molecular Operating Environment (MOE)[34] and—as an additional descriptor—the absolute potency value of the first compound in the MMP.[31] Potency differences between MMPs were used as dependent variables. Regression models were derived for terminal leaf nodes of RF trees that were then used to predict potency differences for test MMPs. For RF generation, the number of trees was set to 400. For all other randomForest parameters, default settings were used.

Performance Evaluation. Balanced accuracy (BA) and F-scores were calculated for classification and the root-mean-square error (RMSE) of AC potency differences for regression. An MMP might have been assigned "compound-out" or "both-out" status, respectively, more than once during cross-validation cycles with varying prediction results, given the stochastic procedure that is steadily reshuffling the subset of kept-out compounds. Therefore, each prediction was included in BA or RMSE calculations after weighing by the inverse of the number of occurrences of the MMP in the given set: Let $E^{\mathrm{set}}_{\mathrm{MMP}}$ be the number of times a given MMP was assigned to the external set ∈ {compound-out, both-out}. In addition, let recorded predictions be $P^{\mathrm{set}}_{\mathrm{MMP}}(k)$, $k = 1...E^{\mathrm{set}}_{\mathrm{MMP}}$. Given the experimental value $P^{\mathrm{expt}}_{\mathrm{MMP}}$, the contribution of a given MMP to the cross-validated RMSE over set is

$$\mathrm{ERR_{MMP}}^2(\mathrm{set}) = \frac{\sum_{k=1}^{E^{\mathrm{set}}_{\mathrm{MMP}}} [P^{\mathrm{set}}_{\mathrm{MMP}}(k) - P^{\mathrm{expt}}_{\mathrm{MMP}}]^2}{E^{\mathrm{set}}_{\mathrm{MMP}}} \tag{1}$$

which results in a global RMSE of:

$$\mathrm{RMSE}(\mathrm{set}) = \sqrt{\frac{\sum_{\mathrm{MMP}=1}^{N_{\mathrm{set}}} \mathrm{ERR_{MMP}}^2(\mathrm{set})}{N_{\mathrm{set}}}} \tag{2}$$

where N_{set} represents the global number of MMPs present at least once in set.

For classification, consider an MMP being predicted A times in its correct class, whereas the remaining $E^{\mathrm{set}}_{\mathrm{MMP}} - A$ prediction attempts were incorrect. Therefore, this MMP will contribute a fractional increment of $f = A/E^{\mathrm{set}}_{\mathrm{MMP}}$ to the count of correct predictions. Thus, if this MMP represents an AC, the true positive (TP) count will be incremented by f, while the false negative (FN) count will be increased by $1 - f$. Otherwise (noncliff), increment f will be added to the true negative count

(TN) and $1 - f$ to false positives (FP). Balanced accuracy (BA), precision, recall, and F-scores were then calculated as follows:

$$\mathrm{BA} = \frac{0.5\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} + \frac{0.5\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}} \quad (4)$$

$$\mathrm{precision}(P) = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \quad (5)$$

$$\mathrm{recall(or\ true\ positive\ rate,\ TPR)} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \quad (6)$$

$$\mathrm{Fscore} = 2 \times \frac{P \times \mathrm{TPR}}{P + \mathrm{TPR}} \quad (7)$$

RESULTS AND DISCUSSION

MMP and Activity Cliff Statistics. As reported in Table 1, the compound data sets used herein yielded large numbers of MMPs, ranging from 366 (for 58 janus kinase inhibitors) to 14 493 (for 2202 factor Xa inhibitors) as well as significant numbers of ACs, ranging from 109 to 1161. In addition, the numbers of MMPs with potency differences within 1 order of magnitude (non-AC MMPs) are reported, ranging from 186 to 10 108, which served as negative training instances. Thus, many more noncliffs (a total of 31 795) than ACs (a total of 4210) were available for model building and predictions, consistent with the observation that ACs are generally rare in sets of active compounds.[1,2] In addition to the requirement to focus predictions on compound pairs, the AC vs noncliff unbalance of the data sets presented another challenge for machine learning.

Performance of Order-Independent Classification Models. F-scores and BA were calculated to evaluate the predictions of classification models. Figure 3 compares model performance for different cross-validation strategies. Furthermore, the results are compared to SVM classification using the special MMP kernel designed for AC predictions, as reported previously.[15] Detailed prediction statistics are reported in Tables S1 and S2 of the Supporting Information. From Figure 3, it is evident that the classification models reached consistently high performance levels with F-scores close to or greater than 0.8 comparable to our previous AC predictions.[15] The performance was slightly lower for the stricter compound-out and both-out strategies than item-out cross-validation, as one might expect.

Given that the MMP data sets contained many more non-ACs than ACs, resulting in imbalanced composition, "balanced accuracy" (see Concepts, Methods, and Materials) was calculated throughout for predictions using classification models and all data sets. In addition, we have also separately monitored retrieval rates for ACs and non-ACs, as reported in Table 2. Technically, the rate of correct prediction of ACs, which count as "true positives" according to our classification mode, equals the recall factor in eq 6, also called "sensitivity", whereas the rate of correct prediction of non-ACs represents the "specificity" (referring to "true negatives"). For all cross-validation strategies, the specificity for predicting non-AC pairs, which dominated the composition of the data sets, was consistently higher than 90%. Given the dominance of non-AC pairs (ACs are generally rare[2]), these accuracy levels were not unexpectedly high for SVM-based classification.[15] However, sensitivity was generally lower, as one also might expect in light of data imbalance (Table 2a–c), and the challenging "compound-out" and "both-out" AXV scenarios had a marked impact on it. Under most challenging cross-validation conditions, sensitivity ranged from 55.0% to 91.5%. Encouragingly, for three data sets, the separately monitored AC retrieval rate was greater than 80% and for three others, it was ~64–66%. It was consistently lowest for kappa opioid receptor ligands (59.7–55.0%) and highest for calpain 2 inhibitors (93.4–91.5%). Item-out calculations yielded F-scores comparable to those obtained for the MMP kernel. Note, however, that MMP kernel calculations were originally reported in the context of a more lenient 10-fold item-out cross-validation scheme, compared to the repeated three-layered scheme used here. Thus, CGR representations of MMPs are well-suited for AC predictions, the practical importance of this result being not as much a gain in terms of statistical performance, but the intrinsic simplicity of the CGR-based approach, alleviating the need to employ specialized kernel functions.

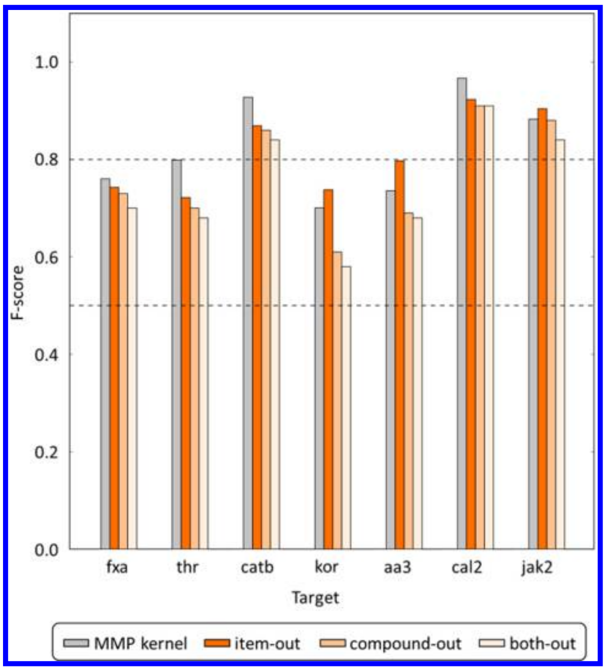


Figure 3. Performance of classification models. The bar chart reports F-scores for the OICG-based classification models (obtained in 3-fold cross-validation) and the, previoulsy reported,[15] MMP kernel-based classification models (obtained in 10-fold cross-validation). The item-out, compound-out, and both-out cross-validations are represented by color-coded bars.

We note that target-dependent variations in the accuracy of AC classifications are expected for at least two reasons. First, target-specific compound sets may have varying degrees of imbalance, a general phenomenon affecting AC predictions; second, activity predictions are in general compound class-dependent and this dependence also applies to ACs that are formed in a compound class- and target-dependent manner.

Performance of Order-Dependent Regression Models. Going beyond AC classification, two strategies were applied to build regression models. Figure 4 reports the performance of regression models including ODCG-based (Figure 4a) and MQSPR descriptor-based models (Figure 4b). Detailed prediction statistics are reported in Table S3 and S4 of the Supporting Information. In the case of regression, the mean

Table 2. Classification Performance of ACs vs non-ACs[a]

(a) Item-out XV

data set	no. ACs	% correct ACs (sensitivity)	no. non-ACs	% correct non-ACs (specificity)
Factor Xa inhibitors	1161	70.69	10108	97.74
Thrombin inhibitors	1103	69.74	6390	95.94
Cathepsin b inhibitors	120	84.40	451	97.44
Kappa opioid receptor ligands	649	59.70	7190	97.17
Adenosine A3 receptor ligands	681	67.39	6752	97.66
Calpain 2 inhibitors	387	93.37	718	95.14
Janus kinase inhibitors	109	92.46	186	92.94

(b) Compound-out XV

data set	no. ACs	% correct ACs (sensitivity)	no. non-ACs	% correct non-ACs (specificity)
Factor Xa inhibitors	1161	69.10	10108	97.65
Thrombin inhibitors	1103	67.87	6390	95.71
Cathepsin b inhibitors	120	82.86	451	97.09
Kappa opioid receptor ligands	649	57.34	7190	97.13
Adenosine A3 receptor ligands	681	65.41	6752	97.51
Calpain 2 inhibitors	387	92.62	718	94.71
Janus kinase inhibitors	109	89.65	186	92.01

(c) Both-out XV

data set	no. ACs	% correct ACs (sensitivity)	no. non-ACs	% correct non-ACs (specificity)
Factor Xa inhibitors	993	66.29	8785	97.57
Thrombin inhibitors	973	64.01	5519	95.61
Cathepsin b inhibitors	105	80.95	391	97.11
Kappa opioid receptor ligands	567	55.00	6250	96.94
Adenosine A3 receptor ligands	607	64.44	5924	97.44
Calpain 2 inhibitors	341	91.45	624	94.30
Janus kinase inhibitors	85	84.20	156	91.48

[a] Reported are the percentage of correctly predicted ACs and non-ACs for OICG-based classification models and item-out XV. "% correct" gives the percentage of correctly classified instances.

RMSE value for the evolved "top" models was calculated as the final performance measure. Figure 4 shows that the RMSE values of the MQSPR-based regression models, ranging from 0.34 to 0.45 for item-out, was much lower compared to ODCG-based models, ranging from 0.54 to 0.68 for item-out. Thus, for regression analysis, predicting the directionality of potency changes encoded by ACs, the MQSPR MMP descriptors were clearly preferred, yielding overall reasonable performance. However, unlike OI classification models (vide supra), regression models witnessed a significant performance decrease in AXV. This may be partly due to the artifact of having considered the two order-dependent (C_1, C_2) and (C_2, C_1) MMPs as independent "items" in item-out XV. Indeed, (C_1, C_2) might have been predicted by a local model featuring the "mirror image" (C_2, C_1) within its training set. Statistically, this happens in two cases out of three, while in one case both (C_1, C_2) and (C_2, C_1) would have been together in the test set. This scenario is impossible with OI-approaches, where it is guaranteed that compounds in a left-out pair would never appear *together* within the local training set. Thus, if regression seems much more vulnerable to the encounter of completely novel test instances than classification, this may be partly due to the fact that "item-out" XV results are overly optimistic.

Even if the very precise estimation of an activity shift is not within reach of these OD-regression models, they nevertheless remain useful. For example, accuracy of prediction of the *sign* of activity change—not accounted for OI-models—is excellent, even in the challenging "both-out" AXV scenario: at worst (factor Xa inhibitors), in only 4% of the situations featuring an activity increase of absolute magnitude above 0.5, the predicted potency difference failed to return the correct sign. Furthermore, important data set-dependent differences were observed. For example, in the case of calpain 2 inhibitors, prediction errors under challenging cross-validation conditions were much larger than for cathepsin b inhibitors or adenosine A3 receptor ligands.

In Figure 4c, the performance of the ODCG- and MQSPR descriptor-based regression models is compared to a previously reported RF modeling approach carried out as a control. The results show that the SVR models generally yielded lower RMSE values than the RF models, an exception being both-out ODCG-based SVR and RF where the RMSE value distribution were comparable. Furthermore, the boxplots reveal that MQSPR-based SVR performed overall better than ODCG-based SVR.

Figure 5 addresses the question of the accuracy of predictions of the particularly "deep" ACs. For this purpose, prediction accuracy (here, item-out cross-validated RMSE) was specifically monitored over a series of increasingly restricted subsets of MMPs corresponding to ACs with larger potency differences than a moving potency difference threshold. This threshold scanned the range between zero (all MMPs, global RMSE reported) to 4.0 (specific RMSE for the few "deepest" ACs reported), with a step size of 0.5. As can be seen, prediction accuracy was significantly reduced with increasing potency differences. However, Figure 5 also emphasizes that very deep ACs were extremely rare. For example, the cathepsin b data set was the second smallest under study. The left-most extreme point reports the RMSE for only 15 ACs with logarithmic potency differences equal to or larger than 3.5. The other problematic case was the kappa opioid receptor data set, which represented a genuine challenge in terms of modelability, also yielding poor classification results. With these two exceptions, reductions in prediction accuracy were not dramatic for MMPs/ACs with increasing potency differences. In general, prediction of rare instances is most challenging for machine learning and typically error-prone.

Effects of the Fragment Control Applicability Domain on Model Performance. Predictive models were derived within the boundaries of training set instances in descriptor space, delineating applicability domains for predictions. Since the underlying descriptor space was defined by ISIDA fragment

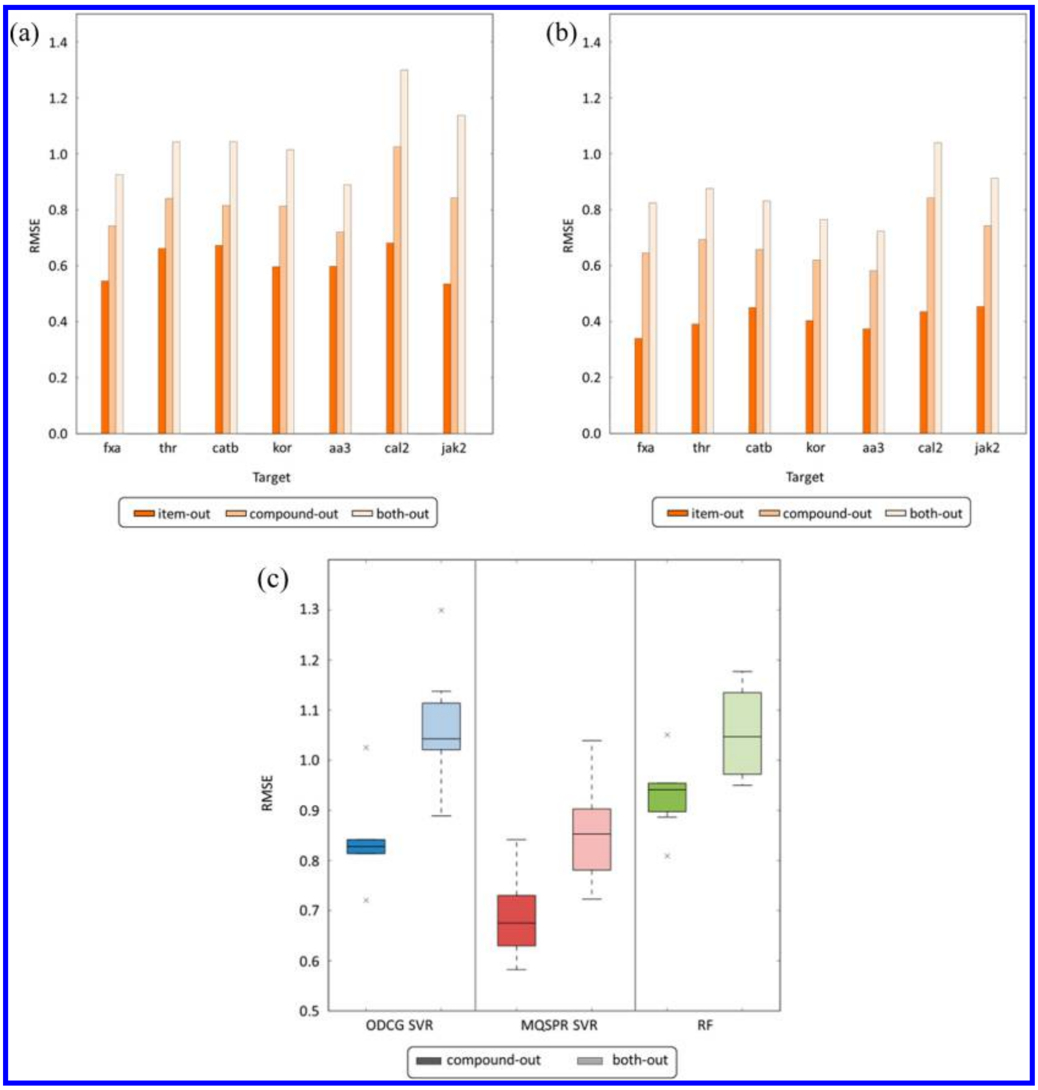


Figure 4. Performance of regression models and control calculations. The bar charts report RMSE values for (a) ODCG-based regression and (b) MQSPR descriptor-based regression models The item-out, compound-out, and both-out cross-validations are represented by color-coded bars. (c) Performance of the ODCG- and MQSPR descriptor-based regression models compared to RF models in box plot representations.

descriptors, items containing fragment counts that do not respect the boundaries observed for training instances are technically "outside" the "bounding box" delimiting the chemical space zone that served for model training. Therefore, one may "trust" only predictions of items inside this bounding box. In order to check whether such selective prediction would actually lead to improved accuracy, the AXV procedure was also employed to monitor, for each predicted MMP, any violations of training set-based bounding boxes. Interestingly, however, this was not the case. Tables 3, 4, and 5 report F-scores and RMSE values, with and without removal of AD violators. For all compound sets, the F-score of the classification models remained either unchanged or displayed only a marginal increase (ranging from 0 to 0.06) when the AD filtering was applied. Similar observations were made for the regression models. Only in one instance, i.e. cathepsin b inhibitors, a notable reduction in prediction errors occurred due the AD criterion. Thus, confinement within the bounding-box tested here was not a major determinant of predictive performance. Robust prediction accuracy was retained outside the AD, i.e. ACs and non-ACs represented by MMPs with previously unobserved novel structural information were well predicted, an encouraging finding for AC classification and regression.

■ CONCLUDING DISCUSSION

Although activity cliffs are extensively studied in medicinal chemistry and chemical informatics, few attempts have thus far been reported to systematically predict them using machine learning. In this study, the condensed graph of reaction formalism, which was originally developed for reaction modeling, was adopted to represent MMP-based ACs as a single graph. In addition, descriptor recombination was applied. Both approaches led to the representation of MMPs/ACs as single descriptor vectors as input for machine learning. ACs were accurately classified using SVM modeling, without the need to employ sophisticated kernels. In addition, SVR regression was successfully applied, for the first time, to predict the directionality of potency changes encoded by MMPs/ACs. The ability to correctly predict these changes has definite potential for virtual screening-type practical applications. For

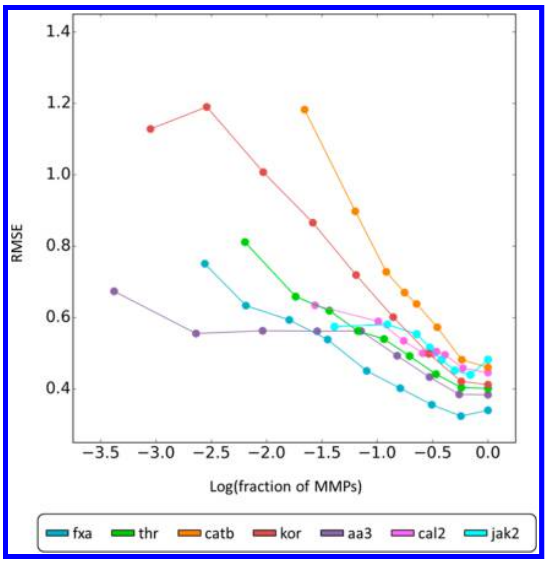


Figure 5. Regression performance for MMP subsets with increasing potency differences. Prediction error (item-out XV RMSE, on Y axis) is monitored with respect to subsets of MMPs having an experimental potency difference (absolute value) exceeding or equal to a moving threshold value scanning the range from 0 to 4.0, with a step of 0.5. The X axis reports (log scale) the relative fraction of MMPs represented by the respective subsets. Every dot on the plot corresponds to a given threshold value. From the right to the left, the threshold value is increasing and MMP subset fraction decreasing. The right-most point at threshold zero represents 100% of the MMPs of a data set and thus reports the global RMSE. For some of the target sets, the left-most point corresponds to a potency difference threshold of 3.5, because there were none, or too few, MMPs with potency differences ≥4.0 for calculation of statistically meaningful RMSE values.

Table 3. Effects of Applicability Domains on Classification Models[a]

	F-score (compound-out)		F-score (both-out)	
data set	without AD	with AD	without AD	with AD
Factor Xa inhibitors	0.73	0.73	0.71	0.71
Thrombin inhibitors	0.70	0.71	0.68	0.69
Cathepsin b inhibitors	0.85	0.90	0.84	0.90
Kappa opioid receptor ligands	0.61	0.61	0.58	0.59
Adenosine A3 receptor ligands	0.69	0.70	0.68	0.69
Calpain 2 inhibitors	0.92	0.92	0.91	0.92
Janus kinase inhibitors	0.88	0.92	0.84	0.91

[a] Reported are F-scores for classification models and different cross-validation schemes in the presence and absence of applicability domains.

Table 4. Effects of Applicability Domain on ODCG-Based Regression Models[a]

	RMSE (compound-out)		RMSE (both-out)	
data set	without AD	with AD	without AD	with AD
Factor Xa inhibitors	0.75	0.73	0.95	0.91
Thrombin inhibitors	0.85	0.83	1.05	1.03
Cathepsin b inhibitors	0.88	0.70	1.14	0.92
Kappa opioid receptor ligands	0.82	0.81	1.02	1.01
Adenosine a3 receptor ligands	0.73	0.72	0.90	0.89
Calpain 2 inhibitors	1.04	0.99	1.36	1.23
Janus kinase inhibitors	0.76	0.75	1.04	0.98

[a] Reported are RMSE values for ODCG-based regression models and different cross-validation schemes in the presence and absence of applicability domain.

Table 5. Effects of Applicability Domain on MQSPR Regression Models[a]

	RMSE (compound-out)		RMSE (both-out)	
data set	without AD	with AD	without AD	with AD
Factor Xa inhibitors	0.64	0.64	0.82	0.82
Thrombin inhibitors	0.69	0.67	0.88	0.8
Cathepsin b inhibitors	0.66	0.52	0.83	0.68
Kappa opioid receptor ligands	0.62	0.61	0.77	0.75
Adenosine a3 receptor ligands	0.58	0.57	0.72	0.73
Calpain 2 inhibitors	0.84	0.74	1.04	0.99
Janus kinase inhibitors	0.74	0.68	0.91	0.79

[a] Reported are RMSE values for MQSPR descriptor-based regression models and different cross-validation schemes in the presence and absence of applicability domain.

example, the compound pair-based approach can be applied to directly search for highly potent compounds, which falls outside the applicability domain of similarity searching. For this purpose, compound pairs are systematically assembled with known weakly potent molecules and test compounds and regression models as derived herein applied to predict the directionality and magnitude of associated potency changes. Predicted large-magnitude ACs then contain test compounds that are candidates for high potency. Although this work was based on innovations concerning the chemoinformatic treatment of MMPs, i.e. their conversion into CGRs and analogy to MQSPR, respectively, there was no need to develop novel, dedicated descriptors. Irrespective of modeling strategies, ISIDA fragment descriptors yielded satisfactory results. Another interesting observation was that binary AC vs non-AC discrimination was essentially possible for MMPs having one or even both compounds never encountered at the model training stage. This means that the approach did not only learn from compound pairs (C_i, C_j) presented at the training stage in order to extrapolate the behavior of so-far not encountered compound combinations, but made it possible to extrapolate to novel structures. By contrast, regression approaches resulted in increased prediction errors when challenged to predict pairs featuring one or two compounds never encountered before. However, this increase might at least in part be linked to the strategic decision to consider MMPs with permuted compound order as two distinct items upon training. Since in item-out XV a given MMP might have served for training the model that predicted its "mirror image", the baseline item-out XV statistics might be too optimistic. This issue was not analyzed in depth here, since the overall quality of the model should always be taken from the most challenging validation experiment, i.e. AXV results. These were sufficiently accurate to foresee very large (>2 orders of magnitude) alterations in activity, and well suited to predict the sign of the activity shift. It is noted that the OI-classification and OD-regression model were completely independent, but could be carried out in a consensus fashion where the prediction of the former would be used to judge the pertinence of the latter; unless the OI-classification would predict the MMP to be an AC, any large potency difference returned by regression should be subject to caution.

Taken together, the results reported herein broaden the methodological spectrum for AC prediction and establish SVR for the prediction of AC-encoded potency alterations. The ability to predict potency changes of large magnitude among structurally analogous compounds should be of interest for a variety of drug design applications.

One of the major conclusions of this work is that QSPR-type modeling of ACs is readily accessible. Technologies such as CGR or plain descriptor recombination strategies, such as used in MQSPR, are versatile enough to support both order-dependent and -independent approaches and complement more complex kernel-based methods for order-independent modeling.

■ ASSOCIATED CONTENT

Ⓢ Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jcim.6b00359.

> Supplementary Methods: calculation details and system settings. Supplementary Tables S1−S4: detailed prediction statistics for different classification and regression models. Supplementary Table S5: ISIDA fragmentation statistics (PDF)

■ AUTHOR INFORMATION

Corresponding Authors

*E-mail: varnek@unistra.fr (A.V.).
*E-mail: bajorath@bit.uni-bonn.de (J.B.).

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

S.K. thanks the Region Alsace for a Ph.D. fellowship.

■ ABBREVIATIONS

AC, activity cliff; AXV, advanced (compound-out and both-out) cross-validation; CGR, condensed graph of reaction; MMP, matched molecular pair; MQSPR, QSPR for nonadditive mixtures; QSPR, quantitative structure−property relationships; OD, order-dependent; OI, order-independent; SAR, structure−activity relationship; SVM, support vector machine; XV, cross-validation

■ REFERENCES

(1) Maggiora, G. M. On Outliers and Activity Cliffs – Why QSAR often Disappoints. *J. Chem. Inf. Model.* **2006**, *46*, 1535−1535.

(2) Stumpfe, D.; Hu, Y.; Dimova, D.; Bajorath, J. Recent Progress in Understanding Activity Cliffs and Their Utility in Medicinal Chemistry. *J. Med. Chem.* **2014**, *57*, 18−28.

(3) Willett, P.; Barnard, J. M.; Downs, G. M. Chemical Similarity Searching. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 983−996.

(4) Stumpfe, D.; Bajorath, J. Similarity Searching. *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **2011**, *1*, 260−282.

(5) Kenny, P. W.; Sadowski, J. Structure Modification in Chemical Databases. In *Chemoinformatics in Drug Discovery*; Opera, T. I., Ed.; Wiley-VCH; Weinheim, Germany, 2004; pp 271−285.

(6) Griffen, E.; Leach, A. G.; Robb, G. R.; Warner, D. J. Matched Molecular Pairs as a Medicinal Chemistry Tool. *J. Med. Chem.* **2011**, *54*, 7739−7750.

(7) Hussain, J.; Rea, C. Computationally Efficient Algorithm to Identify Matched Molecular Pairs (MMPs) in Large Data Sets. *J. Chem. Inf. Model.* **2010**, *50*, 339−348.

(8) Hu, X.; Hu, Y.; Vogt, M.; Stumpfe, D.; Bajorath, J. MMP-Cliffs: Systematic Identification of Activity Cliffs on the Basis of Matched Molecular Pairs. *J. Chem. Inf. Model.* **2012**, *52*, 1138−1145.

(9) Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5−32.

(10) Guha, R. Exploring Uncharted Territories: Predicting Activity Cliffs in Structure-Activity Landscapes. *J. Chem. Inf. Model.* **2012**, *52*, 2181−2191.

(11) Kennedy, J.; Eberhart, R. C. Particle Swarm Optimization. *Proc. IEEE Inter. Conf. Neural Networks* **1995**, 1942−1948.

(12) Namasivayam, V.; Bajorath, J. Searching for Coordinated Activity Cliffs Using Particle Swarm Optimization. *J. Chem. Inf. Model.* **2012**, *52*, 927−934.

(13) Vapnik, V. N. *The Nature of Statistical Learning Theory*, 2nd Ed.; Springer: New York, 2000.

(14) Joachims, T. Making Large-Scale SVM Learning Practical. In *Advances in Kernel Methods—Support Vector Learning*; Schölkopf, B., Burges, C. J. C., Smola, A. J., Eds.; MIT-Press: Cambridge, MA, 1999; pp 169−184.

(15) Heikamp, K.; Hu, X.; Yan, A.; Bajorath, J. Prediction of Activity Cliffs Using Support Vector Machines. *J. Chem. Inf. Model.* **2012**, *52*, 2354−2365.

(16) de la Vega de León, A.; Bajorath, J. Prediction of Compound Potency Changes in Matched Molecular Pairs Using Support Vector Regression. *J. Chem. Inf. Model.* **2014**, *54*, 2654−2663.

(17) Drucker, H.; Burges, C.; et al. Support Vector Regression Machines. *Adv. Neural Inform. Process. Systems* **1997**, *9*, 155−161.

(18) de Luca, A.; Horvath, D.; Marcou, G.; Solov'ev, V.; Varnek, A. Mining Chemical Reactions Using Neighborhood Behavior and Condensed Graphs of Reactions Approaches. *J. Chem. Inf. Model.* **2012**, *52*, 2325−2338.

(19) Hoonakker, F.; Lachiche, N.; Varnek, A.; Wagner, A. Condensed Graph of Reaction: Considering a Chemical Reaction as One Single Pseudo Molecule. *Int. J. Artificial Intelligence Tools.* **2011**, *20*, 253−270.

(20) Oprisiu, I.; Varlamova, E.; Muratov, E.; Artemenko, A.; Marcou, G.; Polishchuk, P.; Kuz'min, V.; Varnek, A. QSPR Approach to Predict Nonadditive Properties of Mixtures. Application to Bubble Point Temperatures of Binary Mixtures of Liquids. *Mol. Inf.* **2012**, *31*, 491−502.

(21) Liu, T.; Lin, Y.; Wen, X.; Jorissen, R. N.; Gilson, M. K. BindingDB: a Web-Accessible Database of Experimentally Determined Protein-Ligand Binding Affinities. *Nucleic Acids Res.* **2007**, *35*, D198−D201.

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

(23) *ChemAxon API*, ChemAxon: Budapest, Hungary, 2016.

(24) Strasbourg web server. http://infochim.u-strasbg.fr/webserv/VSEngine.html (accessed April 7, 2016).

(25) Varnek, A.; Fourches, D.; Horvath, D.; Klimchuk, O.; Gaudin, C.; Vayer, P.; Solov'ev, V.; Hoonakker, F.; Tetko, I. V.; Marcou, G. ISIDA - Platform for virtual screening based on fragment and pharmacophoric descriptors. *Curr. Comput.-Aided Drug Des.* **2008**, *4*, 191−198.

(26) Ruggiu, F.; Marcou, G.; Varnek, A.; Horvath, D. ISIDA Property-Labelled Fragment Descriptors. *Mol. Inf.* **2010**, *29*, 855−868.

(27) Horvath, D.; Brown, J. B.; Marcou, G.; Varnek, A. An Evolutionary Optimizer of libsvm Models. *Challenges.* **2014**, *5*, 450−472.

(28) Chang, C.-C.; Lin, C.-J. LIBSVM: A Library for Support Vector Machines. *ACM Trans. Intell. Syst. Technol.* **2011**, *2*, 1−27.

(29) Muratov, E. M.; Varlamova, E. V.; Artemenko, A. G.; Polishchuk, P. G.; Kuz'min, V. E. Existing and Developing approaches for QSAR Analysis of Mixtures. *Mol. Inf.* **2012**, *31*, 202−221.

(30) Ellison, C. M.; Enoch, S. J.; Cronin, M. T.; Madden, J. C.; Judson, P. Definition of the Applicability Domains of Knowledge-

Based Predictive Toxicology Expert Systems by Using a Structural Fragment-Based Approach. *ATLA, Altern. Lab. Anim.* **2009**, *37*, 533−545.

(31) Beck, J. M.; Springer, C. Quantitative Structure-activity Relationship Models of Chemical Transformations from Matched Pairs Analyses. *J. Chem. Inf. Model.* **2014**, *54*, 1226−1234.

(32) *R: A Language and Environment for Statistical Computing*, version 2.14.1; R Foundation for Statistical Computing: Vienna, Austria, 2008.

(33) Liaw, A.; Wiener, M. Classification and Regression by randomForest. *R News* **2002**, *2*, 18−22.

(34) *Molecular Operating Environment (MOE)*, version 2014.09; Chemical Computing Group Inc.: Canada, 2011.

Supporting Information

In the following, a detailed explanation of descriptor calculation and model generation have been described.

Rendering MMPs as condensed graphs of reaction (CGR)

A CGR captures within a single, connex graph all the relevant information that is present in the classical, multi-partner representation of a chemical process. Furthermore, this strategy is not limited to rendering actually feasible chemical transformations - any "alchemical" transformation can be described in this way, which makes the CGR technology well suited to represent MMPs. Formally, an MMP represents a "transformation" in which a substituent is "broken off" from the core and "replaced" by another. Previous MMP selection and curation work[117] resulted in input files rendering MMPs as a triplet of SMARTS of the core, first, and second substituent "radicals". Free valences in both core C and substituents $S1$, $S2$ were formally represented as bonds to the "Any" query atom "*", meaning that the initial MMP information was given as $(C - *, S1 - *, S2 - *)$, understood to describe the pair of compounds $(C - S1, C - S2)$. A dedicated java tool based on the ChemAxon API[118] for the handling of molecular structures has been written in order to convert above-mentioned SMARTS triplicates into condensed graphs in SDF format. It will create two dynamical bonds between the anchor atom of the core and each connecting atom of $S1$ and $S2$, respectively. By default, the tool operates in "OD" mode: being sensitive to the order of triplicates $(C - *, S1 - *, S2 - *)$ it will conventionally assign the $C - S1$ bond a "forming" status, whilst $C - S2$ is defined as a "breaking" bond. Reversely, upon input of $(C - *, S2 - *, S1 - *)$, the "symmetric" condensed graph $S1-$(breaking bond)$-C-$(forming bond)$-S2$ would be generated. In OI mode, however, both of the dynamical bonds will be given "forming" status, so that the same CG, $S1 -$ (forming bond) $- C -$ (forming bond) $- S2$ will be generated irrespective of ordering of the input triplet. Note that the java tool also assigns mapping numbers (it "marks", as will be stated in current terminology) to the three key atoms in the CG: the core anchor is mapped as #1, the connector in the first and second fragments become #2 and #3, respectively. Also, note that as a "byproduct" of the creation of the CG, the mentioned java tool will also regenerate the two structures of individual compounds

$(C - S1, C - S2)$, in which core and substituent anchors are being marked as well. These individual compounds were subjected to the customary standardization procedure on the basis of the virtual screening database of the Strasbourg web server.[119] As input compounds were already preprocessed, many standardization steps (salt and organometallics removal, size filters, etc.) are not relevant here. However, conversion of nitro/nitroxides to split charge representations, and "basic" ChemAxon aromatization after conversion into the ChemAxon preferred tautomeric form did produce relevant changes in some structures. Note that condensed graphs could not have been standardized in this way, for they would have been rejected as "irregular" molecules, but since the input fragments were already carefully curated, this problem will have no impact on descriptor calculation (vide infra).

Descriptor calculation

All descriptor sets used here were issued from various fragment counting strategies supported by the very flexible ISIDA fragmentor tool. This program allows to control the type of fragments to be counted (sequences or augmented atom fragments), the coloring of the atoms (from default coloring by atomic symbol to pH-dependent pharmacophore or force field typing) to toggle the choice of including or ignoring the nature of the chemical bonds, to specifically focus on dynamical bonds and/or provide a special status to marked atoms. All these options are relevant in the current context. As the fragmentor tool supports so many degrees of freedom, a wealth of different fragmentation schemes were used in a context-dependent manner, in order to provide large pools of candidate descriptor spaces - for both classification and regression problems. Out of these pools of candidate descriptor spaces, the best suited for predictive model building will be singled out during the evolutionary SVM model optimization process.

In the following, only the general principles used to define the proposed context-specific fragmentation scheme pools will be given. Note that all structures, pertaining to all target-specific data sets were first joined together into unique molecular files submitted to fragment counting. In this way, the total number of distinct fragments found in these structures can be monitored, for each of the initially envisaged fragmentation schemes. Since a key parameter in defining fragmentation schemes is setting the boundaries for the smallest and respectively largest fragments to count, upper boundaries producing fragment-rich schemes with more than F (15,000, unless otherwise stated) distinct fragments were discarded. Fragmentation schemes including bond type information are intrinsically more fragment-rich than the ones focusing only on atoms, and will have to stick to smaller upper size boundaries. Upper size (-u option in fragmentor tool) was subject to a scan, steadily increasing it until the threshold of resulting distinct fragment was attained: the most fragment-rich scheme immediately below that threshold was kept.

Order-independent fragmentation schemes

Descriptors for OI-classification models were extracted from ordering-independent condensed graphs (OICG). These contain two dynamical bonds of same type, and the head atoms of these bonds are marked. Since both of these strategies to single out the changing moiety in the MMP are equivalent, using the marked atom flag (-m option in fragmentor tool) for fragmentation would not help to capture any additional information not already highlighted by the presence of the dynamical bond. Fragmentation schemes exploiting the dynamical bond status were thus envisaged, using options to enumerate all the fragments in the OICG (-d 0 option in fragmentor tool), and the alternative choice (-d 1 option in fragmentor tool) to count only the fragments containing at least one dynamical bond. In order to fully exploit dynamical bond information, we privileged fragmentation schemes explicitly accounting for bond orders ("AB" type according to fragmentor nomenclature), which means that fragments C-C-C-C and C-C-(forming bond)-C-C will be considered as distinct descriptor elements (and the former will not count unless the -d 0 option is used). Note that, if the user chooses to ignore bond orders, then above fragments will be fused into one common category CCCC, irrespective of the single, multiple or dynamical nature of the bonds between the carbons. "A" type fragments with option -d set to 0 treat the OICG as a regular molecule. Nonetheless, except for situations when one of the substituents of the MMP is a plain H atom, not explicitly monitored in this approach, the peculiar topology of the core atom with one "valence" more than customarily, may suffice to create a specific "signature" in terms of fragment counts, so that the key information about the CG may not be lost albeit the dynamical nature of the bonds is ignored. Furthermore, when the -d option is set to 1 in an "A" type fragmentation scheme, only fragments containing dynamical bonds count, even though bonding information is not explicit. In this scenario, the "CCCC" count will report the total number of confounded C-(dynamical bond)-C-(dynamical bond)-CC, C-(dynamical bond)-CCC and CC-(dynamical bond)-CC fragments, all while ignoring all the CCCC sequences not shared between the core/substituents. In terms of atom coloring, CG-based descriptors cannot benefit from any elaborate atom typing scheme, such as pharmacophore or force field type assignment.[40] In the context of chemically senseless bond orders and/or core atoms with valence overflow, atom typing tools would invariably fail. Furthermore, the pharmacophore type of the core anchor may actually change within the MMP - a protonable secondary amine (cation, H bond donor) may convert into aniline N (H bond acceptor) or an amide N (featureless). Therefore, CG-based fragmentation schemes exclusively employed element type symbols for atom labels.

Order-dependent fragmentation schemes

Two distinct strategies were used to generate descriptors for OD-regression approaches:

1. Ordering-dependent condensed graphs (ODCG): In this scheme as well both -d 0 and -d 1 option were explored, with atom symbols as labels. Unlike previous case, explicit inclusion of bond information using “AB” type of fragments is mandatory here because the order of substituents may only be read from the dynamical bond orders. “A” type fragments based on ODCGs would be intrinsically order-independent, thus inappropriate. Also it should be noted that marked atoms may acquire special status in the ISIDA fragment counting process (-m 3 option in fragmentor tool). The tool only distinguishes between unmarked and marked atoms, but ignores the actual mapping labels: atoms marked 1, 2 and 3 can be differentiated as being “special” in contrast to all other atoms of the CG, but the ISIDA fragmentor tool would not further distinguish between core and substituent anchors. Thus, the “marked atom” fragmentation strategy, which originally did not encode substituent order was actually toggled in “AB” fragment types to provide an alternative “point of view” to the key atoms (marker atom descriptor spaces include a “m” label in their names). The compulsory inclusion of bond order information and the presence of two distinct types of dynamical bonds mechanically lead to a significant increase, at similar fragment size boundaries, of the number of possible fragments in ODCG descriptors, when compared to OICG. In this particular context, frag-

ment size boundaries were not systematically scanned, but "borrowed" from the AB-type OICG fragmentation schemes that were most successful in classification models. Even with the most conservative estimates for upper fragment sizes, the total number of fragments emerging from ODCGs rapidly exceeded the customary threshold of F=15,000. With ODCGs, fragmentation schemes generating up to F=30,000 distinct fragments were allowed.

2. Descriptors based on MQSPR strategy: Here, the 39 usual fragmentation schemes known, from previous experience[6] to be instrumental in successfully solving various QSPR problems were used as departure basis for the design of candidate descriptor files. These fragmentation schemes were declined in two versions, without (default) and with consideration of the marked atoms (heads of the core-substituent bond, in each molecule). The latter version produced larger fragment numbers, and schemes exceeding the F threshold of 15,000 were discarded. Then, the generated molecular fragment files were used to build up MQSPR descriptor vector $D(C-S1, C-S2)$, by concatenating, for each MMP, the sum and respectively difference of the descriptor vectors of the two participating compounds $C-S1$ and $C-S2$. Formally, considering that the total number of fragments produced by the current scheme is f, then:

$$D_{\mathrm{i}}(C-S1, C-S2) = \begin{cases} D_{\mathrm{i}}(C-S1) + D_{\mathrm{i}}(C-S2) & \forall\, 1 \leq i \leq f \\ D_{\mathrm{i}}(C-S1) - D_{\mathrm{i}}(C-S2) & \forall\, f \leq i \leq 2f \end{cases} \tag{5.1}$$

where, $D_{\mathrm{i}}(C-S1, C-S2)$ is the concatenated value of the sum and difference part of the MQSPR descriptor vector for fragment i. The first part of the vector $D(C-S1, C-S2)$, i.e., the summation is obviously invariant to the ordering of the compounds in the MMP, whereas the difference section values change sign upon inversion of the order of compounds in the pair.

Model building procedures

In the former order-dependent case (regression), property-descriptor matrices systematically enumerated both formal pairs corresponding to swapped compound ordering: $\Delta(C-S1, C-S2) :: D(C-S1, C-S2)$ and $-\Delta(C-S1, C-S2) :: D(C-S2, C-S1)$, respectively. In classification calculations, MMPs with intermediate potency difference scores were discarded, in order to avoid misclassification artefacts due to such borderline items. The evolutionary optimizer of libsvm-based SVM models was then used to search for optimally cross-validating predictive models, for each of the seven target-specific data sets and each modeling strategy. This was envisaged in order to avoid premature convergence of the evolutionary procedure towards one family of descriptors, i.e., failure to sufficiently sample the other alternative. Preliminary simulations showed that the default 3,000 evolutionary generations would not be needed for the discovery of good predictive models in this context. This number was changed to 400. The model fitness criterion is based on 3-fold "item-out" cross-validation, which was repeated 12 times, after data reshuffling. Note that, in regression models, this leave-item-out cross-validation ignores the special relationships between the pair of items corresponding to swapped compound orders. Leaving both such pairs out, simultaneously, versus having them split up between training and left-out tiers are both possible, in randomly occurring scenarios. For each of the 12 cross-validation trials, the cross-validated correlation coefficient (Q^2) or balanced accuracy are determined for regression or classification models, respectively. The mean of these 12 values, penalized by twice their standard deviation, represents the chosen fitness function, i.e., the predictive power of the SVM model at given parameters (descriptor set, kernel type, cost, gamma, etc.).

A posteriori, controlled model cross-validation and applicability domain assessment

After completion of above-mentioned evolutionary simulations, and for each of the explored target-specific compound set, a list of best models, were ordered by their fitness scores. From these lists, entries based on the same descriptor spaces already encountered in even fitter solutions were discarded, and the top four to ten models (or all qualifying) - each best in its descriptor space - were kept for in-depth "compound-out" and "both-out" cross-validation strategies, and monitoring of the impact of applicability domain filters on these cross-validation results. The following stochastic cross-validation procedure was assigned and a flowchart showing the steps is given in **Figure 2** of the publication.

1. The list of all compounds involved in the MMPs of the current target-specific series is established. This will serve as a pool for random sampling of left-out compounds, without replacement. Let the total number of compounds

involved be C.

2. At the current iteration, an approximate number of C/5 compounds are randomly picked from the pool. Since this picking is without replacement, once the number of molecules still in the pool drops below C/5, all will be picked and the empty pool will be reinitialized to the full set of C compounds, in view of future iterations.

3. For each MMP of the series, if none of the pair members figure amongst the currently picked compounds, then the MMP is moved to the training set. If, by contrast, only one of the pair members figures amongst the picked, this MMP will be dispatched to the "compound-out" test set. Eventually, if the MMP is a combination of two picked members, it will be registered within the "both-out" test set. Note that, unlike in the point-out cross-validation at model building, both entries associated to a swapped compound pair will be always be assigned to the same set.

4. With the current training set as defined above, the SVM model - using the operational parameters found by the evolutionary algorithm - is then retrained.

5. The applicability domain of this model is defined by a "bounding box", a hyper-parallelepiped in descriptor space, bounded by the minimal and respectively maximal descriptor values encountered within training set compounds. Note that if any given fragment i is absent in training compounds, then the bounding box limits with respect to axis i will be (0,0) - any other item with a non-zero value for D_{i} would be counted as a violator of the bounding box.

6. The trained model is applied, independently, to "compound-out" and "both-out" test sets, and predictions for the therein present MMPs are redirected to specific output files, in view of separate statistical evaluation of the prediction proficiency of both scenarios.

7. The compliance of test set MMPs with the currently established bounding box is checked, by counting the number of violations (number of descriptor elements falling outside the allotted ranges). This number of violations is stored, in order to enable determination of specific statistical parameters for the prediction cases that were AD-compliant, where "compliance" can be defined in

terms of tolerated number of violations. Here, statistics will be reported at zero violations (strict compliance) and any number of violations (which is the baseline prediction quality, in absence of AD considerations).

Iterations continue until, ideally, all the MMPs have acquired, at least once, the "both-out" validation status. Since the stochastic procedure might take very much time until this desideratum is fulfilled, it was decided to stop it after ten complete "cycles", where a cycle was defined to begin every time the empty pool of compounds needed to be reset to the initial list. Eventually, predictions associated with compound-out and both-out configurations, and annotated by AD compliance, were processed in order to calculate the associated RMS Errors (RMSE) of potency differences (regression) and balanced accuracy and F-scores (classification).

Data set	**Model**	**F-score (item-out)**	**Mean F-score (item-out)**
Factor Xa inhibitors	Model 1	0.77	0.74
	Model 2	0.77	
	Model 3	0.76	
	Model 4	0.74	
	Model 5	0.71	
	Model 6	0.72	
	Model 7	0.70	
Thrombin inhibitors	Model 1	0.75	0.72
	Model 2	0.76	
	Model 3	0.71	
	Model 4	0.71	
	Model 5	0.71	
	Model 6	0.69	
	Model 7	0.70	

Cathepsin b inhibitors	Model 1	0.89	0.87
	Model 2	0.88	
	Model 3	0.89	
	Model 4	0.86	
	Model 5	0.88	
	Model 6	0.84	
	Model 7	0.81	
Kappa opoid receptor ligands	Model 1	0.79	0.74
	Model 2	0.78	
	Model 3	0.77	
	Model 4	0.72	
	Model 5	0.71	
	Model 6	0.69	
	Model 7	0.67	
Adenosine A3 receptor ligands	Model 1	0.82	0.79
	Model 2	0.81	
	Model 3	0.81	
	Model 4	0.79	
	Model 5	0.77	
	Model 6	0.77	
	Model 7	0.77	
Calpain 2 inhibitors	Model 1	0.96	0.92
	Model 2	0.94	
	Model 3	0.92	
	Model 4	0.92	
	Model 5	0.89	
	Model 6	0.90	
	Model 7	0.90	
Janus kinase inhibitors	Model 1	0.96	0.90
	Model 2	0.94	
	Model 3	0.92	
	Model 4	0.92	
	Model 5	0.89	
	Model 6	0.90	
	Model 7	0.90	

Table 5.1: F-score of OICG based classification models. The table reports the F-scores of individual OICG-based classification models with item-out XV and the consensus F-score over all models.

Data set	**Model**	**Balanced accuracy (item-out)**	**Mean balanced accuracy (item-out)**
Factor Xa inhibitors	Model 1	0.86	0.84
	Model 2	0.85	
	Model 3	0.85	
	Model 4	0.83	
	Model 5	0.83	
	Model 6	0.82	
	Model 7	0.82	
Thrombin inhibitors	Model 1	0.84	0.82
	Model 2	0.84	
	Model 3	0.83	
	Model 4	0.82	
	Model 5	0.82	
	Model 6	0.81	
	Model 7	0.81	
Cathepsin b inhibitors	Model 1	0.93	0.90
	Model 2	0.92	
	Model 3	0.92	
	Model 4	0.91	
	Model 5	0.91	
	Model 6	0.89	
	Model 7	0.85	
Kappa opoid receptor ligands	Model 1	0.82	0.78
	Model 2	0.81	
	Model 3	0.80	
	Model 4	0.77	
	Model 5	0.77	
	Model 6	0.75	
	Model 7	0.74	
Adenosine A3 receptor ligands	Model 1	0.84	0.82
	Model 2	0.84	
	Model 3	0.83	
	Model 4	0.81	
	Model 5	0.81	
	Model 6	0.81	
	Model 7	0.81	

Calpain 2 inhibitors	Model 1	0.97	0.94
	Model 2	0.95	
	Model 3	0.94	
	Model 4	0.94	
	Model 5	0.92	
	Model 6	0.92	
	Model 7	0.92	
Janus kinase inhibitors	Model 1	0.94	0.92
	Model 2	0.94	
	Model 3	0.93	
	Model 4	0.92	
	Model 5	0.91	
	Model 6	0.92	
	Model 7	0.91	

Table 5.2: Balanced accuracy of OICG based classification models. The table reports the balanced accuracy of individual OICG-based classification models with item-out XV and the consensus balanced accuracy over all models.

Data set	**Model**	**RMSE (item-out)**	**Mean RMSE (item-out)**
Factor Xa inhibitors	Model 1	0.52	0.55
	Model 2	0.53	
	Model 3	0.53	
	Model 4	0.56	
	Model 5	0.57	
Thrombin inhibitors	Model 1	0.61	0.66
	Model 2	0.62	
	Model 3	0.64	
	Model 4	0.70	
	Model 5	0.72	
Cathepsin b inhibitors	Model 1	0.57	0.67
	Model 2	0.61	
	Model 3	0.63	
	Model 4	0.65	
	Model 5	0.67	
	Model 6	0.73	
	Model 7	0.81	

Kappa opoid receptor ligands	Model 1	0.57	0.59
	Model 2	0.58	
	Model 3	0.59	
	Model 4	0.63	
Adenosine A3 receptor ligands	Model 1	0.55	0.59
	Model 2	0.60	
	Model 3	0.60	
	Model 4	0.62	
Calpain 2 inhibitors	Model 1	0.61	0.68
	Model 2	0.64	
	Model 3	0.65	
	Model 4	0.64	
	Model 5	0.73	
	Model 6	0.72	
	Model 7	0.75	
Janus kinase inhibitors	Model 1	0.40	0.54
	Model 2	0.39	
	Model 3	0.44	
	Model 4	0.45	
	Model 5	0.51	
	Model 6	0.65	
	Model 7	0.88	

Table 5.3: RMSE of ODCG based regression models. The table reports the RMSE values of individual ODCG-based regression models with item-out XV and the consensus RMSE over all models.

Summary

This study further extends the current spectrum of methodologies for prediction of activity cliffs. Compound pairs forming MMPs were encoded either as condensed graphs or by descriptor recombination (MQSPR). ISIDA fragment count descriptors were calculated for each of these representations to serve as an input for the support vector machine models. In addition, two different modeling strategies were explored in this work. Categorical (order-independent) strategy were used to build SVC models for predicting activity cliff/non-cliff status of a given compound pair. On the other hand, SVR models based on quantitative (order-dependent) modeling strategy was introduced as the first approach for predicting directional potency changes in MMPs. In addition, advanced cross-validation schemes were applied to evaluate model performance under stringent conditions of leaving either one or both of the test set MMP compounds out of the training phase. As expected, the performance of both SVC and SVR models were affected in advanced cross-validation, with SVR models showing relatively much larger decrease in their performance compared to SVC. Nevertheless, even under strict cross-validation conditions our predictive models yielded satisfactory results, indicating their ability to extrapolate to novel structures not observed during training stage.

My major contributions to this work have been generating GTM-based classification models for selected data sets and evaluating model performance for activity cliffs and non-cliffs under item-, compound-, and both-out cross validation scenarios.

Biological activity is undoubtedly an important compound property governing successful interaction of drugs and targets. However, activity alone is not the sole determinant of the efficacy of resulting candidate compound and several other properties need to be simultaneously optimized alongside activity during compound optimization. In the next chapter, we analyze ionization state of bioactive compounds, one of the cardinal parameters affecting pharmacological profile of compounds.

Chapter 6

Target-based Analysis of Ionization States of Bioactive Compounds

Introduction

The ionization state of a compound influences several important physicochemical properties, ultimately affecting their biological activity and drug-action. Majority of currently available drugs are ionized under physiological conditions, with an estimation of 75% of weak bases and 20% of weak acids among them. As such, the charge state of drug molecules is altered by the variation in pH levels in their surroundings. This in turn affects their properties *in vivo*, such as drug permeability, that determines the possibility of their diffusion across different cellular membranes. Moreover, the charge state of compounds at varying pH conditions greatly affects the absorption, distribution, metabolism, excretion and toxicity (ADMET) of drugs. Because of their impact on various biopharmaceutical and pharmacokinetic properties of compounds, considerable amount interest has been placed in exploring ionizaion profiles of drugs and bioactive compounds. Comparison between the ionization state of drugs and screening hits have revealed that ionizable groups such as carboxylic acids and aliphatic amines are significantly underrepresented in screening sets. Furthermore, notable differences in charge state distribution across different target superfamilies as well as among individual families within a given superfamily.

The current study presents an extensive analysis of ionization state profiles of publicly available bioactive compounds from high-confidence activity sets. The prime focus of this work is to investigate relationships between compound's ionization states, structural features and potency. MMP formalism was used for the first time to determine ionization states in chemical neighborhood of bioactive compounds, to understand the effect of structural transformations on charge state of compounds. Ionization state distribution was monitored globally and among highly potent compounds. Further, charge state preferences of compounds active against specific target families were studied.

MedChemComm

CONCISE ARTICLE

Cite this: *Med. Chem. Commun.*, 2015, **6**, 1030

Target-based analysis of ionization states of bioactive compounds

Shilva Kayastha,† Antonio de la Vega de León,† Dilyana Dimova and Jürgen Bajorath*

Received 6th February 2015, Accepted 17th April 2015

DOI: 10.1039/c5md00051c

www.rsc.org/medchemcomm

A systematic analysis of ionization states of current bioactive compounds is presented. Ionization states were related to biological activities on the basis of high-confidence activity data. The majority of bioactive compounds were found to be basic or neutral under physiological conditions. In addition, chemical neighborhoods of active compounds frequently contained analogs with different ionization states that were activity-conservative. However, a variety of targets were identified that displayed clear preferences for specific ionization states in compounds active against them. In this context, notable differences in the distribution of ionization states were detected for compounds active against different target superfamilies. Furthermore, under physiological pH, differences in ionization states of active compounds were tolerated by many targets. However, in a number of instances, ionization states of highly and weakly potent compounds active against the same target were found to be distinct, providing guidelines for compound design and optimization.

Introduction

The charge state of small molecules is a major determinant of biological activity and drug action.[1–5] It has been estimated that the majority of drugs are partly ionized under physiological conditions.[1] A convincing perspective has also been provided on the critical role compound ionization states play at different stages of pharmaceutical development.[3] Importantly, differences in the pH in various cellular compartments or extracellular environments can modulate ionization states of active compounds, alter their properties *in vivo*, and affect pharmacological profiles. A refined charge state profile of oral drugs indicated that nearly 80% of them contained ionizable groups, while only ~12% were neutral.[3] In addition, acid/base profiles of drugs directed against major target classes (including proteases, kinases, G protein coupled receptors, and various ion channels) were studied and notable differences between these profiles were identified as well as differences between individual target families comprising a given class.[4] Furthermore, the ionization states of drugs and screening compounds were compared. It was found that drugs contained a much higher proportion of both carboxylic acid groups and aliphatic amines than compounds from various sources available for screening, indicating that many compounds in screening collections might lack relevance for drug discovery, given the prevalent charge states of drugs.[4]

In a recent extensive analysis of publicly available compound data,[5] acidic and basic bioactive compounds and drugs were compared and the influence of ionization states on a variety of calculated or observed physico-chemical and pharmacological properties was studied. For this purpose, acids and bases were classified as compounds that were proton donors and acceptors, respectively, and at least 50% ionized under physiological pH of 7.4 (calculated using the Henderson–Hasselbalch equation[6]). Major conclusions from this work included that weak bases containing *N*-heterocycles are frequent among drugs and that their physico-chemical and pharmacological properties are by and large tolerable, that strongly basic compounds should best be avoided due to unfavourable properties, and that acids are under-represented in drugs but should merit further consideration.[5] In addition to their thorough analysis of drugs, Charifson and Walters also analyzed the activity distribution of bioactive compounds with different ionization states across cell-based assays and of compounds tested in at least 20 assays. It was found that acidic compounds were generally less active in cellular assays than compounds with other charge states and that basic compounds were overall less selective than acidic or neutral ones.[5]

Herein we also report a large-scale analysis of bioactive compounds with respect to ionization states, albeit with different focal points. Our analysis primarily focuses on

Department of Life Science Informatics, Bonn-Aachen International Center for Information Technology, Rheinische Friedrich-Wilhelms-Universität Bonn, Dahlmannstr. 2, D-53113 Bonn, Germany. E-mail: bajorath@bit.uni-bonn.de; Fax: +49 228 2699 341; Tel: +49 228 2699 306

† The contributions of these authors should be considered equal.

relationships between compound ionization states, structural similarity, and potency and exclusively uses high-confidence activity data.

Methods

Small molecules can generally be classified according to ionization states as bases, acids, neutral compounds, or zwitterionic molecules.[1,5] The dissociation constant (K_a) is an equilibrium constant determining ionization states. Commonly used is the logarithmic form of the dissociation constant (pK_a), defined as the negative decadic logarithm of K_a ($-\log_{10} K_a$). To account for acidic and basic properties of small molecules, two different pK_a types are considered including the acidic pK_a (A_pK_a) and the basic pK_a (B_pK_a). Following this distinction, A_pK_a is defined as the pK_a for the most acidic group in a given molecule whereas B_pK_a is defined as the pK_a for the most basic group. For all compounds analyzed herein, calculated values of A_pK_a and B_pK_a were extracted from the ChEMBL database[7] (version 19).

Compounds were assigned to four ionization state classes (IS-classes) including basic, acidic, neutral, and zwitterionic compounds on the basis of A_pK_a and B_pK_a values relative to the physiological pH of 7.4, following the approach of Charifson and Walters.[5] Accordingly, compounds with an acidic or basic group were classified as acids or bases, respectively, if they were more than 50% ionized at pH 7.4. In addition, compounds containing acidic and basic groups were classified as acids if the acidic group was more than 50% ionized and the basic group less than 50%, as bases if the basic group was more than 50% ionized and the acidic group less than 50%, and as zwitterionic compounds if both groups were ionized more than 50%. Furthermore, compounds were classified as neutral if acidic and/or basic groups were both ionized less than 50% under physiological pH. If A_pK_a and B_pK_a values were not available for a compound, it was not assigned (NA).

From ChEMBL (version 19), compounds active against human targets at the highest confidence level (confidence score 9) were extracted for which assay-independent equilibrium constants (K_i values) were available as potency measurements. Compounds with multiple measurements for the same target were only considered if all values fell within the same order of magnitude. Then the geometric mean of these was calculated as the final potency annotation. If multiple stereoisomers of a compound with potency within one order of magnitude were available, the compound was retained. All qualifying compounds were organized in individual activity classes (target sets). A total of 719 K_i-based target sets were obtained comprising 80 776 compounds.

To assess structural relationships between active compounds, matched molecular pairs (MMPs)[8] were calculated. MMPs consist of pairs of compounds that are only distinguished by a structural change at a single site (chemical transformation).[8,9] Size restrictions were introduced to limit transformations to small structural modifications.[10] Accordingly, the size (number of heavy atoms) of the shared MMP core had to be at least twice the size of each of the exchanged substructures. In addition, the size of each transformation fragment was limited to a maximum of 13 heavy atoms and the difference between the exchanged fragments to eight heavy atoms.[10] For each target set, transformation size-restricted MMPs were systematically calculated using an in-house implementation of the algorithm by Hussain and Rea[8] utilizing the *OEChem* toolkit.[11] MMPs involving NA compounds were omitted from further analysis. Furthermore, target sets yielding fewer than 50 MMPs were excluded. A total of 338 419 MMPs were obtained that exclusively involved a total of 66 871 IS-class compounds from 290 different target sets.

For each classified compound, its chemical neighborhood was determined by combining all of its MMP partners (structural analogs) within a target set and three neighborhood categories (CATs) were defined as follows: (I) all neighbors, (II) only a subset of neighbors, or (III) none of the neighbors belonged to the same IS-class as the reference compound.

Results and discussion

Compound ionization state class distribution

Fig. 1 reports the IS-class assignment for all qualifying bioactive compounds with available high-confidence activity data (only 8.4% of all compounds could not be assigned to one of the four IS-classes, due to missing pK_a values). Consistent with previous findings that many drugs are weak bases under physiological conditions, we also determined that bases were prevalent among bioactive compounds (39.2%). Interestingly, however, a comparable proportion of bioactive compounds was neutral (38.6%), regardless of their activity. Considering the entire potency range, only 3.5% and 10.3% of active compounds were zwitterionic and acidic, respectively. The observed global distribution over IS-classes was essentially mirrored by a subset set of 39 783 compounds with a potency of at least 100 nM (with relative class deviations $<2\%$).

Chemical neighborhood analysis

We then systematically explored the chemical neighborhoods of compounds in all IS-classes through MMP calculations.

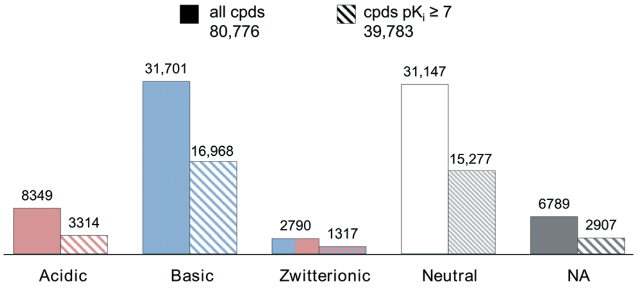


Fig. 1 Ionization state class distribution. Reported is the class distribution for all 80 776 qualifying compounds (solid bars) and a subset of 39 783 compounds with a potency of at least 100 nM or higher (striped bars).

The majority of MMPs (86.4%) were formed between compounds belonging to the same IS-class. Hence, most structural analogs of classified compounds had conserved ionization states. However, many compounds had at least one or a few structural analogs belonging to a different IS-class. We found that 68.8% of all neighborhoods consisted of compounds with conserved ionization states, while 28.7% of the neighborhoods contained one or more compounds belonging to a different IS-class than the reference molecule. In addition, in 2.5% of the neighborhoods, all compounds belonged to IS-classes different from the reference molecule. Hence, about one third of all neighborhoods were heterogeneous in their IS-class composition. However, these frequently occurring differences in ionization states were activity-conservative. Fig. 2 shows exemplary compound neighborhoods of different composition.

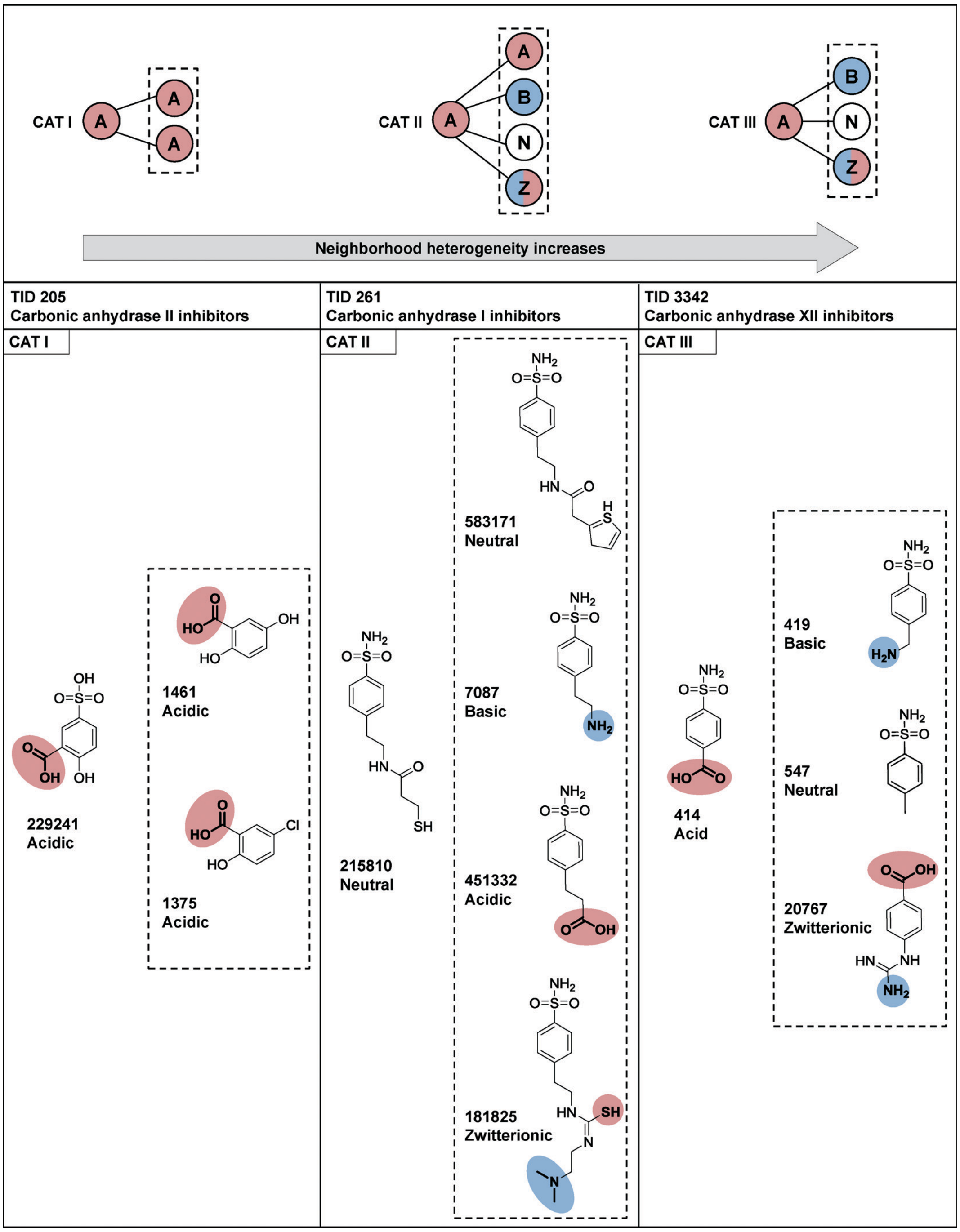


Fig. 2 Chemical neighborhoods. The figure shows the IS-class composition (red, acidic; blue, basic; white, neutral; dual colored, zwitterionic) of exemplary chemical neighborhoods of category (CAT) I–III with inhibitors of different carbonic anhydrase isoforms (TIDs report ChEMBL target set IDs). Functional groups (acidic, red; basic, blue) ionized at physiological pH are depicted in bold and highlighted. ChEMBL compound ID and IS-classes are given.

Ionization state class distribution over activity classes and target superfamilies

We next determined the distribution of IS-classes over target sets. Target sets with fewer than 10 compounds or more than 20% unclassified (NA) compounds were excluded from this analysis. In nearly 90% of 351 qualifying target sets, more than half of the compounds belonged to the same IS-class and in 40%, more than 80% belonged to the same class. Hence, although compound neighborhoods were frequently found to be heterogeneous in their ionic state composition, as discussed above, many target sets displayed a strong ionization state preference. In most cases, basic or neutral compounds dominated. This can also be seen in Table 1 that reports the top 20 target sets (comprising at least 200 compounds) having the highest percentage of compounds belonging to the same IS-class. The ranking contains many different G protein coupled receptors (GPCRs), but also transporters and proteases. In addition to basic compounds, strong preferences for neutral (*e.g.*, vanilloid receptor ligands) and acidic compounds (*e.g.*, prostaglandin D2 receptor 2 ligands) were also observed.

Target sets were also organized into superfamilies and the IS-class distribution of their ligands was determined, as reported in Fig. 3. Clear trends were observed. For example, 50% of available enzyme inhibitors and 36% of membrane receptor ligands were neutral. While enzyme inhibitors displayed a balanced distribution of acidic (15%) and basic (20%) compounds, membrane receptor ligands showed a notable preference for basic (46%) over acidic compounds (8%). Furthermore, 67% of all compounds active against transporters were bases. Moreover, 50% and 26% of ion channel ligands were basic and neutral compounds, respectively, whereas only 3% of them were acids. Thus, there were marked differences in ionization state preferences for compounds active against different target superfamilies.

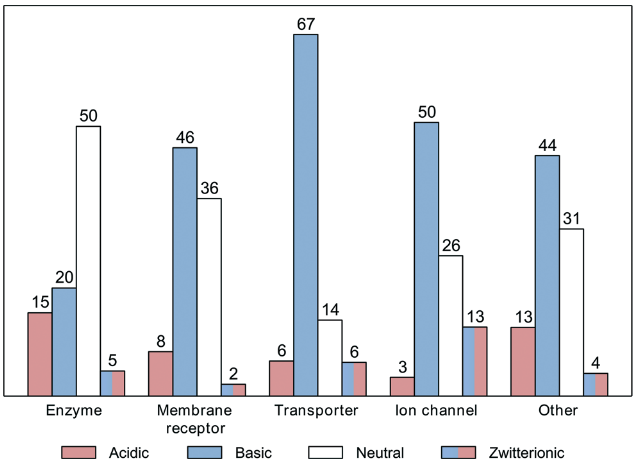


Fig. 3 Compound IS-class distribution over target superfamilies. Reported is the IS-class distribution (red, acidic; blue, basic; white, neutral; dual colored, zwitterionic) of ligands of four target superfamilies (plus "Other"). For each superfamily, the percentage of active compounds belonging to each category is given. Compounds classified as NA are not shown.

Potency range distribution of ionization state classes

Finally, the potency range distribution of IS-classes was studied in detail for all target sets. Although there were no significant differences between the global IS-class distributions of all bioactive compounds and a subset of highly potent compounds, as reported above (and shown in Fig. 1), we detected 57 target sets with notable differences in IS-class distributions between weakly ($pK_i \leq 6$; WP) and highly potent ($pK_i \geq 7$; HP) compounds. These 57 target sets primarily belonged

Table 1 Target sets with ionization state class dominance[a]

Target ID	Target name	# Cpds	Dominant IS-class
5071	Prostaglandin D2 receptor 2	468	99% acidic
4794	Vanilloid receptor	253	97% neutral
259	Melanocortin receptor 4	1217	92% basic
264	Histamine H3 receptor	2023	92% basic
1898	Serotonin 1b (5-HT1b) receptor	364	92% basic
335	Protein-tyrosine phosphatase 1B	243	91% acid
344	Melanin-concentrating hormone receptor 1	846	90% basic
4644	Melanocortin receptor 3	350	90% basic
4608	Melanocortin receptor 5	268	88% basic
1983	Serotonin 1d (5-HT1d) receptor	359	87% basic
1800	Corticotropin releasing factor receptor 1	473	84% neutral
222	Norepinephrine transporter	1010	84% basic
232	Alpha-1b adrenergic receptor	290	84% basic
228	Serotonin transporter	1337	83% basic
2492	Neuronal acetylcholine receptor protein alpha-7 subunit	253	83% basic
238	Dopamine transporter	867	81% basic
3798	Calcitonin gene-related peptide type 1 receptor	349	81% neutral
1916	Alpha-2c adrenergic receptor	295	80% basic
2954	Cathepsin S	375	80% neutral
210	Beta-2 adrenergic receptor	241	80% basic

[a] The top 20 target sets (with ChEMBL IDs) with highest percentages of compounds belonging to the same IS-class are reported.

to three superfamilies (enzymes, membrane receptors and transporters). The majority of the targets belonged to enzyme inhibitors (28) followed by membrane receptors (19). Only two targets were transporters. The IS class distribution of HP and WP compounds in these target sets displayed significant differences. In 27 enzyme inhibitor sets, more than 80% of the HP compounds were zwitterionic (and less than 20% of WP compounds were zwitterionic), whereas for the majority of membrane receptors, the percentage of HP zwitterionic compounds was less than 20%. The structures of the HP and WP compound sets for specific IS-classes and superfamilies were mostly distinct, because the overlap in scaffolds[12] between these sets of compounds rarely exceeded 5%. In addition, the sets of HP and WP compounds had high intra-set diversity, because each scaffold represented on average only one to two different compounds. In Fig. 4, representative examples for these 57 target sets are shown. Fig. 4A reports the IS-classes of neurokinin 2 receptor antagonists. Among the highly potent ligands, there was a clear preference for basic over neutral compounds, whereas the trend was reversed for weakly potent compounds where neutral species were found to dominate. In Fig. 4B, a notable enrichment of basic compounds among weakly potent urokinase-type plasminogen activators is observed. In this case, highly potent compounds had different ionization states (which was rather unusual). Furthermore, Fig. 4C shows a reversal in the distribution of acidic and neutral compounds among highly and weakly potent inhibitors of inosine-5′ monophosphate dehydrogenase 2, corresponding to observations made for basic and neutral compounds in Fig. 4A. Furthermore, Fig. 4D shows that basic compounds were frequently observed among weakly potent coagulation factor XI

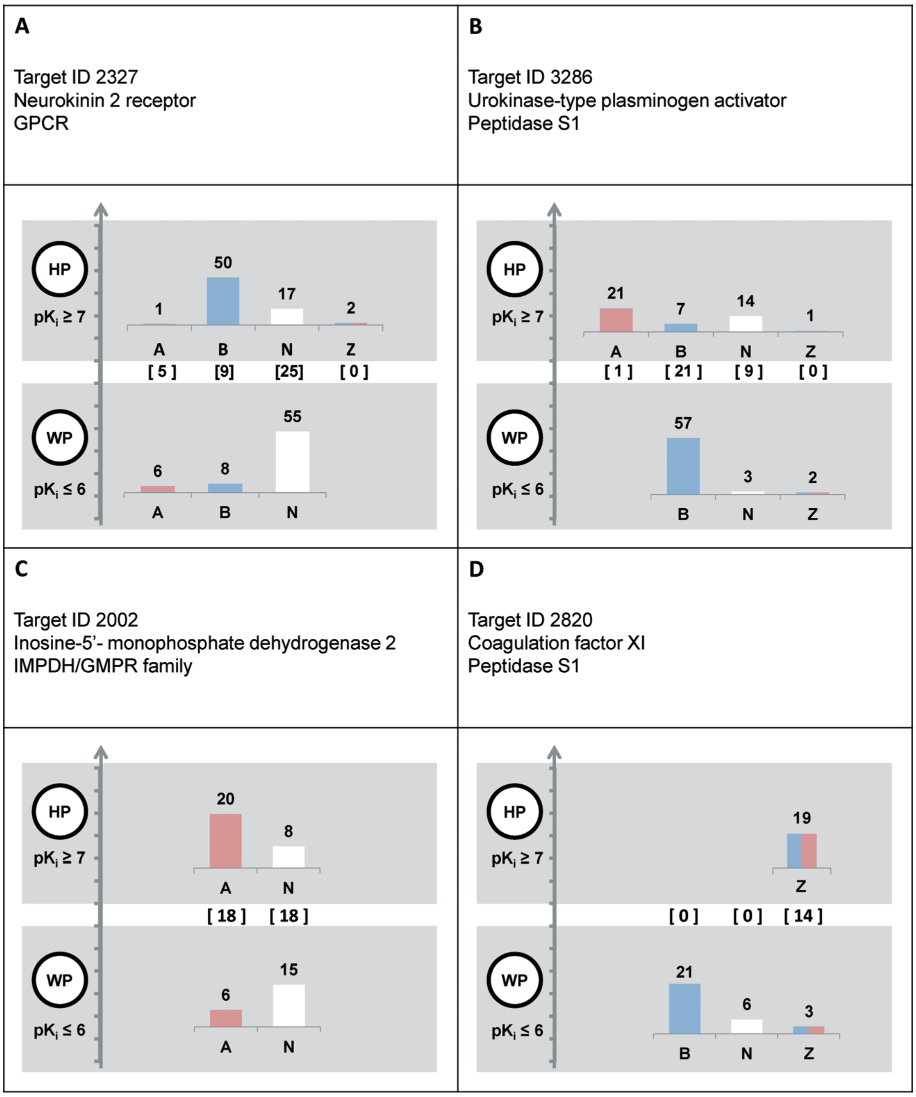


Fig. 4 IS-class changes over potency ranges. For four exemplary target sets (with targets belonging to different families), the IS-class distribution (red, acidic; blue, basic; white, neutral; dual colored, zwitterionic) for highly potent (HP, pK_i ≥ 7) and weakly potent (WP, pK_i ≤ 6) compounds is shown. The number of compounds in different IS-classes falling into the intermediate potency interval is also given. Target IDs: (A) 2327, (B) 3286, (C) 2002, and (D) 2820.

inhibitors, whereas highly potent inhibitors were zwitterionic, without an exception.

Conclusions

We have carried out a large-scale analysis of calculated ionization states in bioactive compounds and their distribution across different targets and families that complements and further extends previous investigations. Ionization states in chemical neighborhoods of bioactive compounds were determined across different potency ranges, setting our analysis apart from previous studies. Furthermore, different from earlier studies that strongly (but not exclusively) focused on drugs, we comprehensively analyzed currently available spectrum of bioactive compounds and exclusively based our analysis on carefully selected high-confidence activity data. Our results reveal the presence of an uneven global distribution of ionization states across the bioactive compounds, the majority of which were basic or neutral under physiological conditions. Individual target sets were found to display significant differences in preferred ionization states. Similar observations were made for different target superfamilies. Systematic MMP analysis revealed that changes in ionization states frequently occurred among structural analogs. Moreover, potency range-dependent differences in the distribution of ionization states were detected in a variety of target sets. We found that ionization states of highly potent compounds were often different from weakly potent ones. In a number of cases, weakly potent compounds were predominantly basic, while different ionization states were observed among highly potent ones. In other instances, weakly potent compounds were mostly neutral, whereas highly potent compounds were charged. Thus, for a variety of targets, preferred ionization states characteristic of highly potent compounds can be identified. The presence of preferred ionization states in highly potent compounds for different targets provides valuable guidelines for compound design and optimization.

Acknowledgements

We are grateful to OpenEye Scientific Software, Inc., for the free academic license of the OpenEye Toolkits. We thank Norbert Furtmann for helpful discussions.

References

1 D. T. Manallack, *Perspect. Med. Chem.*, 2007, **1**, 25–38.

2 M. P. Gleeson, *J. Med. Chem.*, 2008, **51**, 817–834.

3 D. T. Manallack, R. J. Prankerd, E. Yuriev, T. I. Oprea and D. K. Chalmers, *Chem. Soc. Rev.*, 2013, **42**, 485–496.

4 D. T. Manallack, R. J. Prankerd, G. C. Nassta, O. Ursu, T. I. Oprea and D. K. Chalmers, *ChemMedChem*, 2013, **8**, 242–255.

5 P. S. Charifson and W. P. Walters, *J. Med. Chem.*, 2014, **57**, 9701–9717.

6 H. N. Po and N. M. Senozan, *J. Chem. Educ.*, 2001, **78**, 1499–1503.

7 A. Gaulton, L. J. Bellis, A. P. Bento, J. Chambers, M. Davies, A. Hersey, Y. Light, S. McGlinchey, D. Michalovich, B. Al-Lazikani and J. P. Overington, *Nucleic Acids Res.*, 2012, **40**, D1100–1107.

8 E. Griffen, A. G. Leach, G. R. Robb and D. J. Warner, *J. Med. Chem.*, 2011, **54**, 7739–7750.

9 J. Hussain and C. Rea, *J. Chem. Inf. Model.*, 2010, **50**, 339–348.

10 X. Hu, Y. Hu, M. Vogt, D. Stumpfe and J. Bajorath, *J. Chem. Inf. Model.*, 2012, **52**, 1138–1145.

11 OpenEye Scientific Software Inc: Santa Fe, NM.

12 G. W. Bemis and M. A. Murcko, *J. Med. Chem.*, 1996, **39**, 2887–2893.

Summary

A comprehensive analysis of ionizatoin states of currently available bioactive compounds from one of the largest public repositories has been reported. Ionization state of compounds was determined based on two different types of calculated dissociation constant values: acidic pKa and basic pKa, defined as the pKa for the most acidic and the most basic group in the molecule, respectively. Compounds were classified into one of the four ionization state classes: acidic, basic, neutral and zwitterionic, based on their acidic and basic pKa values. The analysis was carried out on a total of 80,776 compounds, meeting high-confidence activity criteria, obtained from 719 targets in ChEMBL. Transformation-size restricted MMPs were systematically calculated for each compound set to account for structural relationships between compounds of different ionization state classes. The chemical neighborhood of each classified compound was determined by considering all of its MMP partners within a target set. Three neighborhood categories were defined if all neighbors (CAT I), only a subset of neighbors (CAT II) or none of the neighbors (CAT III) belonged to the same ionization state class as the reference compound.

Our analysis revealed that the proportion of basic and neural compounds were comparable, and in total, they covered nearly 80% of all bioactive compounds. The chemical neighborhood analysis showed that one third of all neighborhoods had heterogeneous ionization states, i.e., the charge state of the reference and the neighborhood compounds differed. For majority of target classes, compounds showed strong preference towards basic or neutral ionization state. In addition, target superfamilies also displayed significant difference in charge states distribution of the compounds. Detailed analysis of potency distribution of ionization state classes indicated marked differences between weakly and highly potent compounds with respect to their ionization states.

My major contributions to this work include analyzing global and target-based distribution of ionization states of compounds and MMP-based chemical neighborhood analysis of compounds belonging to same or different IS-classes.

As previously outlined, compound optimization is a multifactorial process, and ionization state is just one of the several factors affecting the “drug-likeliness” of a given compound. In the next chapter, computational optimization of multiple drug-relevant properties of compounds have been carried out, followed by visualization of prioritized solutions corresponding to drug-like subspaces of multi-dimensional property space.

Chapter 7

Visualization of Multi-property Landscapes for Compound Selection and Optimization

Introduction

SAR is undoubtedly the most explored theme in medicinal chemistry and chemoinformatics. Although, traditional SAR analysis is done in a case-by-case basis the surge in the amount of compound data requires SAR analysis to be carried out on a large-scale in diverse data sets. Large-scale SAR analysis is thus usually complemented by visualization techniques such as activity landscapes. Several 2D and 3D activity landscapes have been extensively discussed in the *Chapter 1.* The fundamental characteristic of most of the activity landscapes is that only biological activity is considered as the sole feature of compounds. However, compound optimization is dependent on several other biologically relevant properties, which need to be optimized simultaneously. Navigation and visualization of compound distribution in multi-dimensional property spaces are usually difficult, and even more challenging is to identify drug-like subspaces in them. Computational multi-property compound optimization tasks generally result in more than one optimal solutions. It is not straightforward to differentiate and hence choose the globally best combination of the given set of objectives. Moreover, it is often difficult to find a chemically intuitive rationale to the proposed solution of a multi-property optimization problem. This study introduces a method to visualize multi-property landscapes and complement multi-objective optimization by allowing graphical analysis of resulting solutions. Furthermore, the method can be used for comparative analysis of numerically equivalent solutions and prioritize them for identifying compounds with drug-relevant properties for further optimization.

J Comput Aided Mol Des
DOI 10.1007/s10822-015-9862-3

Visualization of multi-property landscapes for compound selection and optimization

Antonio de la Vega de León[1] · Shilva Kayastha[1] · Dilyana Dimova[1] · Thomas Schultz[2] · Jürgen Bajorath[1]

Received: 16 June 2015 / Accepted: 27 July 2015

Abstract Compound optimization generally requires considering multiple properties in concert and reaching a balance between them. Computationally, this process can be supported by multi-objective optimization methods that produce numerical solutions to an optimization task. Since a variety of comparable multi-property solutions are usually obtained further prioritization is required. However, the underlying multi-dimensional property spaces are typically complex and difficult to rationalize. Herein, an approach is introduced to visualize multi-property landscapes by adapting the concepts of star and parallel coordinates from computer graphics. The visualization method is designed to complement multi-objective compound optimization. We show that visualization makes it possible to further distinguish between numerically equivalent optimization solutions and helps to select drug-like compounds from multi-dimensional property spaces. The methodology is intuitive, applicable to a wide range of chemical optimization problems, and made freely available to the scientific community.

Keywords Compound optimization · Activity landscapes · Structure–property relationships · Multi-objective optimization · Multi-property landscapes · Visualization

Introduction

The exploration of structure–activity relationships (SARs) in large and structurally heterogeneous compound data sets is strongly supported by SAR visualization methods [1]. The concept of activity landscapes (ALs) [2] provides integrated views of compound similarity and activity relationships and has been applied for SAR visualization [1, 2]. Several approaches to the design of two- (2D) and three-dimensional (3D) ALs have been introduced that typically consider activity as the sole compound property. Exemplary 2D AL designs include simple "structure–activity similarity (SAS) maps" [3] that plot structural similarity against activity similarity on the basis of pairwise comparisons of data set compounds and, in addition, various network representations. For example, the "network-like similarity graph" (NSG) [4] has been an original network-based AL design in which nodes represent compounds and edges pairwise (fingerprint) similarity relationships. Nodes in NSGs are annotated with potency and numerical SAR score information. Another more recent design has been "intuitive networks for structure–activity relationship analysis" (inSARa) [5] in which reduced graphs of active compounds are used to determine their maximum common substructures (MCSs). These MCSs are then represented as nodes that are connected by edges indicating hierarchical MCS relationships. Original compounds are then assigned to corresponding MCSs and represented as a second node category, i.e., compound nodes colored by potency. MCS-based visualization methods have

Antonio de la Vega de León and Shilva Kayastha have contributed equally to this work.

✉ Jürgen Bajorath
bajorath@bit.uni-bonn.de

[1] Department of Life Science Informatics, B-IT, LIMES Program Unit Chemical Biology and Medicinal Chemistry, Rheinische Friedrich-Wilhelms-Universität, Dahlmannstr. 2, 53113 Bonn, Germany

[2] Institute of Computer Science II, Computer Graphics, Rheinische Friedrich-Wilhelms-Universität, Friedrich-Ebert-Allee 144, 53113 Bonn, Germany

Published online: 02 August 2015

also been introduced to organize individual compound series and elucidate SAR patterns [6–8]. In addition to network representations, tree-like structures have been designed to graphically organize compound series and study SAR trends in chemical neighborhoods [8, 9]. Several network- [7] or tree-like [8, 9] visualizations can be rationalized as local 2D ALs because they predominantly or exclusively focus on individual compound series (rather than structurally heterogeneous data sets).

Going beyond global or local 2D representations, the design of 3D ALs can be approached in different ways. Given a 2D representation of structural similarity relationships, an intuitive way of generating a 3D AL is adding a biological response surface as a third dimension. This typically requires extrapolation of a hypersurface from sparsely distributed compound activity values, which has been accomplished by adapting the kriging method from geostatistics [10]. An alternative approach to 3D AL design is subjecting a high-dimensional chemical descriptor space to dimension reduction to obtain a 3D view, as exemplified by the ligand induced structure–activity relationship display (LiSARD) [11]. Compound positions in this space can then be annotated with activity information.

Chemical space visualization is not confined to AL views. Rather, different visualization techniques have also been introduced to generalize chemical space display including, for example, similarity-based compound networks [12] and molecular layout algorithms [13] for smaller data sets, projections from high-dimensional descriptors spaces based on principal component analysis for large (or very large) data sets [14, 15], and generative topographic mapping (GTM) [16]. GTM was designed to project from high-dimensional feature spaces onto latent 2D space representations in which points (nodes) correspond to normal probability distributions derived from the original data space that determine the mapping of compounds to the latent space. As such, GTM does not represent an AL view as conventionally defined.

Returning to the AL concept, we emphasize two of its cardinal features: firstly, it is activity-centric (i.e., activity is considered as a single structure-related property); secondly, it is descriptive in nature (i.e., ALs are used to analyze SARs but not predict active compounds). Both of these features limit the applicability of AL representations for compound optimization, which typically is a multi-objective task. During iterative optimization, multiple biologically relevant compound properties are considered in combination with activity, focusing on the key question, which compound(s) to make next [17]. In the practice of medicinal chemistry, this process is predominantly driven by chemical experience and intuition, although it can also be supported by computational means. In computational chemistry, multi-property optimization is typically attempted using evolutionary algorithms [18–20] or property-weighted objective functions [20], often in combination with Pareto ranking [19, 20] of numerical solutions. These multi-objective methods usually produce reasonable numerical solutions of optimization tasks but are not expected to find the globally best solution. Multi-objective optimization typically produces a variety of comparable solutions and it is often difficult to further differentiate between them and rationalize characteristic features in multi-dimensional property space.

Herein, we introduce an approach to visualize multi-property landscapes, further extending the AL concept, and graphically analyze solutions of property-weighted objective functions. The methodology makes it possible to further differentiate between numerically equivalent optimization solutions and prioritize them for specific tasks by viewing them in a multi-dimensional data set context.

Materials and methods

Compound data selection

In order to model compound optimization processes, data sets were assembled that consisted of two types of compounds active against the same target: bioactive compounds from medicinal chemistry sources and approved drugs. Bioactive compounds were extracted from ChEMBL [21] (version 20). Only compounds with reported direct interactions (i.e., target relationship type "D") against human targets at the highest assay confidence level (i.e., confidence score 9) and precisely defined equilibrium constants (K_i values) were considered. Compounds with multiple K_i measurements for the same target were retained if all reported values fell within the same order of magnitude. In this case, the arithmetic mean was calculated as the final potency annotation. Approved small molecule drugs with specific target annotations were assembled from DrugBank [22] (version 4.1). To ensure that potency information was available for all drugs and bioactive compounds considered in the analysis, only drugs were retained for which high-confidence activity measurements were available in ChEMBL. All qualifying compounds and drugs with activity against the same target were organized into target-based compound sets. Each target set was required to contain at least 100 bioactive compounds and at least 10 approved drugs. Table 1 summarizes the composition of six target sets satisfying the above criteria assembled for our analysis.

Multi-dimensional property space

A multi-dimensional property space was generated using 14 descriptors accounting for different molecular properties relevant for chemical optimization, as summarized in

Table 1 Data sets combining bioactive compounds and approved drugs

Target ID	Target name	Bioactive CPDs	Drugs
231	Histamine H1 receptor	572	25
1867	Alpha-2a adrenergic receptor	453	23
210	Beta-2 adrenergic receptor	355	19
2035	Muscarinic acetylcholine receptor M5	282	14
4302	P-glycoprotein 1	242	49
4605	Small intestine oligopeptide transporter	181	14

For the six target-based data sets, the ChEMBL target ID, number of bioactive compounds (CPDs), and approved drugs are reported

Table 2. Properties represented by 13 calculated descriptors included, among others, hydrophobic and aromatic character, molecular complexity, hydrogen bonding potential, charge, and surface properties. In addition, compound potency (pK_i; negative decadic logarithm of the equilibrium constant) was used as a descriptor. Experimental pK_i values for data set compounds were taken from ChEMBL (version 20). The descriptor a_ringR (fraction of ring atoms in a molecule) was calculated with the aid of the OpenEye toolkit [23] and the remaining 12 descriptors were calculated using the Molecular Operating Environment (MOE) [24]. This 14-dimensional feature space was designed as a reference space for exemplary multi-property optimization. The feature set selected for our proof-of-concept investigation can of course be replaced by any other number of calculated descriptors and/or experimentally determined properties, depending on the specific optimization tasks.

Property space projection and optimization

Compound subsets with preferred feature value combinations were selected from multi-dimensional feature space. Therefore, compound distributions in 14-dimensional feature space were projected onto a one-dimensional space. A projection of the data was obtained by multiplying an $n \times p$ data matrix, $\boldsymbol{X}$, with n sample points in p dimensions, with a $p \times d$ projection matrix, $\boldsymbol{A}$ (here with $p = 14$ and $d = 1$). Accordingly, the projection of compound i was given by the formula: $val_i = \sum_{j=1}^{p} w_j v_j$, where v_j (from $\boldsymbol{X}$) was the value for descriptor j and w_j (from $\boldsymbol{A}$) the weight given to descriptor j [25]. The value of this projection was used as the *multi-objective function* (MOF) value for numerical optimization of a compound subset selection.

Values of the 13 numerical descriptors were scaled relative to the observed pK_i range to ensure that no descriptors numerically dominated the value distributions.

Table 2 Descriptors

No.	Name	Definition	Property	Unit
1	a_acc	Number of hydrogen bond acceptors	Hydrogen bonding	Integer
2	a_aroR	Fraction of aromatic ring atoms	Aromaticity	Percentage
3	a_don	Number of hydrogen bond donor atoms	Hydrogen bonding	Integer
4	a_ringR	Fraction of ring atoms	Molecular complexity	Percentage
5	b_rotR	Fraction of rotatable bonds	Flexibility	Percentage
6	chiral_u	Number of chiral centers	Stereochemistry	Integer
7	Fcharge	Sum of formal charges	Charge	Integer
8	logP(o/w)	Log of octanol/water partition coefficient	Hydrophobicity	Log unit
9	logS	Log of aqueous solubility	Solubility	Log (mol/L)
10	PEOE_VSA_FHYD	Fractional hydrophobic van der Waals surface area	Surface property	Percentage
11	PEOE_VSA_FPNEG	Fractional negative polar van der Waals surface area	Surface property	Percentage
12	PEOE_VSA_ FPPOS	Fractional positive polar van der Waals surface area	Surface property	Percentage
13	Pot	Potency (pK_i)	Activity	Log (M)
14	Weight	Molecular weight	Molecular size	Da

The set of 14 descriptors used for feature space generation is listed and defined

Optimization was guided by maximizing the MOF value. Therefore, a systematic search was performed using four different weight values for each descriptor {−1.0, −0.33, 0.33, 1.0}. All 4^{14} (~270 millions) possible projections were systematically explored. The weighting scheme chosen for our analysis can be easily exchanged for different properties and optimization tasks. The search procedure is not dependent on a specific methodology or strategy. Descriptor weights can be obtained using alternative approaches including, among others, regression techniques. If the number of features becomes too large for an exhaustive search, stochastic search strategies can also be applied.

Compounds were ranked based on their MOF value and the top 20 compounds were analyzed. Projections were prioritized based on the number of approved drugs within the top 20 ranking. In prioritized set of projections, MOF value corresponded to our drug-likeness model of compounds meaning that compounds with higher MOF values had properties similar to approved drugs. Thus, projections with a significant enrichment of drugs among top-ranked compounds were considered to originate from *drug-like subspaces* representing favorable multi-feature combinations. Our current analysis scheme is focused on the exploration of drug-like subspaces for the generation of which reference sets of known drugs are essential. However, compound reference sets with other characteristic properties of interest can be used for mapping and derivation of descriptor weights.

Visualization of projections

For the visualization of individual projections, the *Star Coordinate* (STC) [26] representation was adopted from computer science. STC is a multi-dimensional visualization technique that arranges coordinates in predefined positions sharing the same origin at the center. The position of a compound in the STC visualization was dependent on the position of each coordinate (descriptor) and the values of the compound for each coordinate.

More formally, the position of compound i in the STC visualization was given by the formula: $\vec{i} = \sum_{j=1}^{p} v_j \vec{d}_j$, where $\vec{d}_j$ represented the position of descriptor j and v_j the value for descriptor j. The position of descriptor j was calculated as follows: its weight obtained from MOF optimization provided the y-axis value. Along the x-axis, all descriptors were ordered lexicographically and given incremental values between −1 and 1 to distribute them evenly. Figure 1a shows a schematic STC visualization for an individual compound. For a given projection, the STC visualization provides a 2D representation of the data set distribution in multi-dimensional property space. STC for multi-property space display was implemented in-house in Java based upon the JUNG library [27].

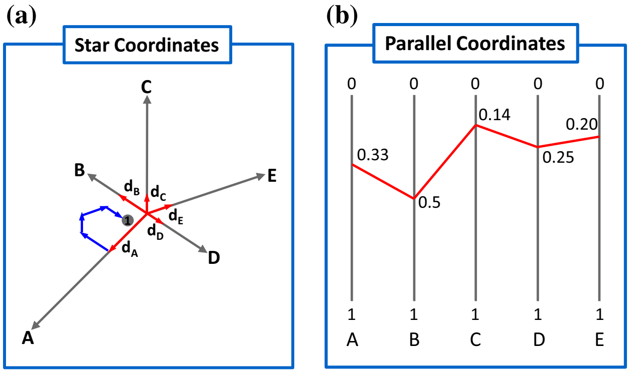


Fig. 1 Star and parallel coordinates. **a** A schematic STC representation for a single compound (*gray dot*) and five descriptors (*A–E*). *Gray arrows* represent descriptor vectors forming the star coordinate. *Red arrows* (d_A to d_E) represent weighted vectors obtained by multiplying the descriptor value of the compound with the corresponding vector. The position of the compound is determined by the sum of all weighted vectors (indicated by *blue arrows* for d_B to d_E). **b** An exemplary PAC plot for the same compound. Descriptors (*A–E*) are assigned to *parallel horizontal lines*. The *red line* traces the descriptor values of the compound

STC visualization was complemented by the *Parallel Coordinate* (PAC) [28] representation, another multi-dimensional visualization technique from computer science that organizes features (descriptors) on parallel axes. Each axis represents all possible values for a descriptor, ranging from the minimum (top of the axis) to the maximum value (bottom). Compounds are then represented as lines that traverse all descriptor axes at positions corresponding to the value for each descriptor. Figure 1b shows an exemplary PAC representation. The molecular PAC representation was also implemented in-house in Java. STC visualizations of projections were generated to further differentiate numerically comparable optimization solutions and view subsets of top-ranked compounds in the context of global data distributions from multi-dimensional feature space.

For comparison, principal component analysis (PCA) of unweighted and weighted descriptor spaces was carried out using R [29] and the first and second principal components (PCs) were used to generate conventional PC plots. Because these plots generate a two-dimensional view of multi-dimensional data that maximize the original variance, they are often used to represent high-dimensional spaces. However, their primary goal is the generation of an uncorrelated view with maximum variance and hence the visualization might not be chemically informative.

Results and discussion

Methodological principles

The simultaneous consideration of multiple properties beyond potency is a requirement of compound optimization in medicinal chemistry. Therefore, the activity-centric AL concept, which is useful for SAR exploration, might be further extended to rationalize multi-property landscapes. Analyzing multi-dimensional property spaces generally is a complicated task, which is typically addressed using dimensionality reduction. The basic idea underlying the methodology introduced herein was to visualize compound distributions in multi-property space in which numerical optimization is carried out. Multi-property optimization carried out in the context of our analysis was guided by the use of approved drugs as internal standards. Compound rankings based upon projections with a significant enrichment of drugs at top ranked positions were thought to originate from drug-like subspaces in multi-dimensional property space. Thus, highly ranked data set compounds had property combinations comparable to drugs and were thus considered preferred candidates for selection and further optimization efforts. A known conundrum of numerical multi-objective optimization is that typically a variety of high-scoring solutions are obtained that are difficult to distinguish. Therefore, it was attempted to visualize compound distributions underlying best projections to analyze rankings within the data set context and further differentiate them. These visualizations were designed to provide a detailed view of multi-property landscapes, as discussed in the following.

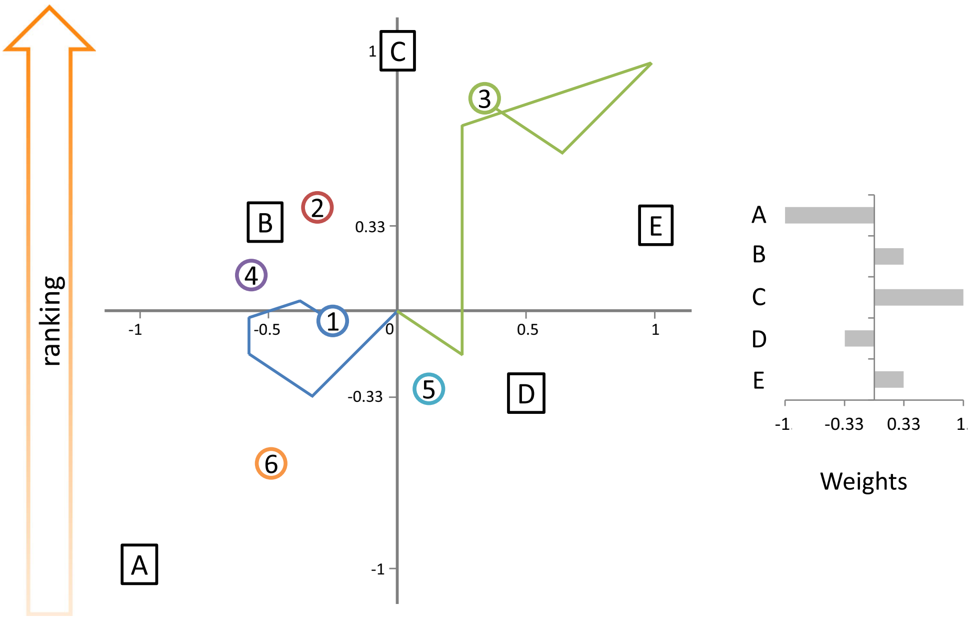
Multi-property landscape display

The STC representation provides the core visualization of multi-property space. Figure 2 illustrates how an STC view is obtained for a model compound set from descriptor weights and multi-dimensional coordinates. For a given projection, the STC visualization provides a 2D view of the underlying compound distributions in multi-dimensional property space. Figure 3a shows an exemplary STC visualization for an actual compound data set and a given projection. In addition, Fig. 3b shows the corresponding PAC view and Fig. 3c the top five compounds from the ranking. The five compounds have similar chemical structure. Hence, MOF value ranking likely includes a similarity-property principle component. At the top of the STC view, the drug-like subspace is delineated by the subset of highly-ranked compounds including 13 drugs, with the majority of compounds being clearly separated from the prioritized subspace (Fig. 3a). Similarity relationships between compounds in STC views were substantially different from those in high-dimensional space (on average 21.5 % nearest neighbors overlap). Comparable average overlap values (12.8–25.2 %) were obtained for other STC views shown in Fig. 5b, d. The corresponding PAC representation reveals which descriptor contributions dominate the projection (Fig. 3b). For some descriptors, values of highly ranked compounds significantly differed (e.g., a_acc, logP(o/w)), whereas their values were narrowly confined in other cases (e.g., a_don, a_ringR, b_rotR). Moreover, largely distinct value ranges of a few descriptors were observed for highly ranked molecules compared to many other bioactive compounds (e.g., a_ringR, b_rotR), which strongly contributed to the separation. Thus, the PAC representation complements the STC visualization by identifying property settings that distinguish compounds in drug-like subspaces from others and evaluating relationships between descriptor settings. Thus, PAC representations can be used to study feature correlation patterns. For example, the line traces in Fig. 3b reveal a negative correlation between the a_ringR and b_rotR descriptors. Finally, PAC also provides a visual representation of the original high-dimensional space, as it displays all descriptor values for each compound. Therefore, the PAC representation is independent of specific projections and helpful to analyze the STC view.

Multi-property optimization

We next carried out a systematic multi-property optimization as a basis for practical applications of the newly introduced visualization approach. The set of chemically intuitive features selected for our conceptual investigation can be replaced by any other calculated or experimentally determined compound characteristics relevant for optimization tasks. For the multi-objective function containing our 14-descriptor set with four possible weights per descriptor, a systematic search of all possible projections from multi-dimensional space was carried out. Each projection yielded a MOF value for any bioactive compound and drug based on which a ranking was generated. More than 270 million weight combinations were analyzed and prioritized based on the number of drugs in the top 20 compound ranking. For all data sets, drug enrichment was only detected in a small subset of possible weight combinations, as shown in Fig. 4. Hence, delineation of subspaces populated with drugs required very specific multi-parameter settings, as one should expect. Nonetheless, for the different data sets, there were between 20 and ~500 projections that yielded maximum

Fig. 2 Star coordinate representation of multi-dimensional compound data. For a model data set comprising six compounds with five different properties, the generation of an STC view is illustrated. Descriptor positions and compound *1* correspond to Fig. 1a. Descriptor positions resulted from lexicographical ordering along the *horizontal axis* combined with weight settings for a given projection (with a descriptor weight combination shown in the *inset*). Compound positions resulted from matrix calculations shown at the *bottom* and summation of descriptor contributions (pathway calculations). For two exemplary compounds, *1* and *3*, pathways are traced. Compound rank positions increase along the *vertical axis*



	A	B	C	D	E
1	0.33	0.50	0.14	0.25	0.20
2	0.88	0.93	0.97	0.80	0.53
3	0.35	0.61	0.89	0.50	0.74
4	0.85	0.88	0.64	0.35	0.54
5	0.60	0.01	0.21	0.30	0.38
6	0.95	0.16	0.38	0.50	0.29

Descriptor values

x

	x	y
A	-1	-1
B	-0.50	0.33
C	0	1
D	0.50	-0.33
E	1	0.33

Descriptor position

=

	x	y
1	-0.26	-0.04
2	-0.42	0.31
3	0.34	0.82
4	-0.58	0.14
5	-0.08	-0.36
6	-0.49	-0.59

Compound position

drug enrichment (between nine and 18 drugs for the different sets), as also shown in Fig. 4. Thus, these projections represented equivalent numerical optimization solutions. The corresponding compound rankings covered most drugs in the data sets (43–95 %; on average 70 %) but only a small fraction of bioactive compounds (3–16 %; on average 8 %) mapping to drug-like subspaces. Furthermore, many projections producing maximum drug enrichment had very similar weight combinations. However, projections with very different combinations (descriptor contributions) were also found. Therefore, solutions with maximal drug enrichment having similar or distinct weight combinations were further analyzed through visualization. The successful delineation of specific drug-like subspaces for all data sets indicated that the search procedure took compound similarity relationship implicitly into account.

Visualization of projections and comparison of compound distributions

A large number of STC representations were generated for different data set projections. Figure 5 shows exemplary comparisons. In Fig. 5a, two projections with distinct descriptor weight combinations are shown for beta-2 adrenergic receptor ligands that produced large drug enrichment (and shared 11 of 13 drugs in their top 20 rankings). Figure 5b compares the STC representations of these projections. The compound distributions differed significantly for these two projections representing numerically equivalent optimization solutions. This might be expected because distinct weight combinations characterized these projections. Although both projections displayed significant drug enrichment, projection 1 clearly separated top ranked compounds from others and also

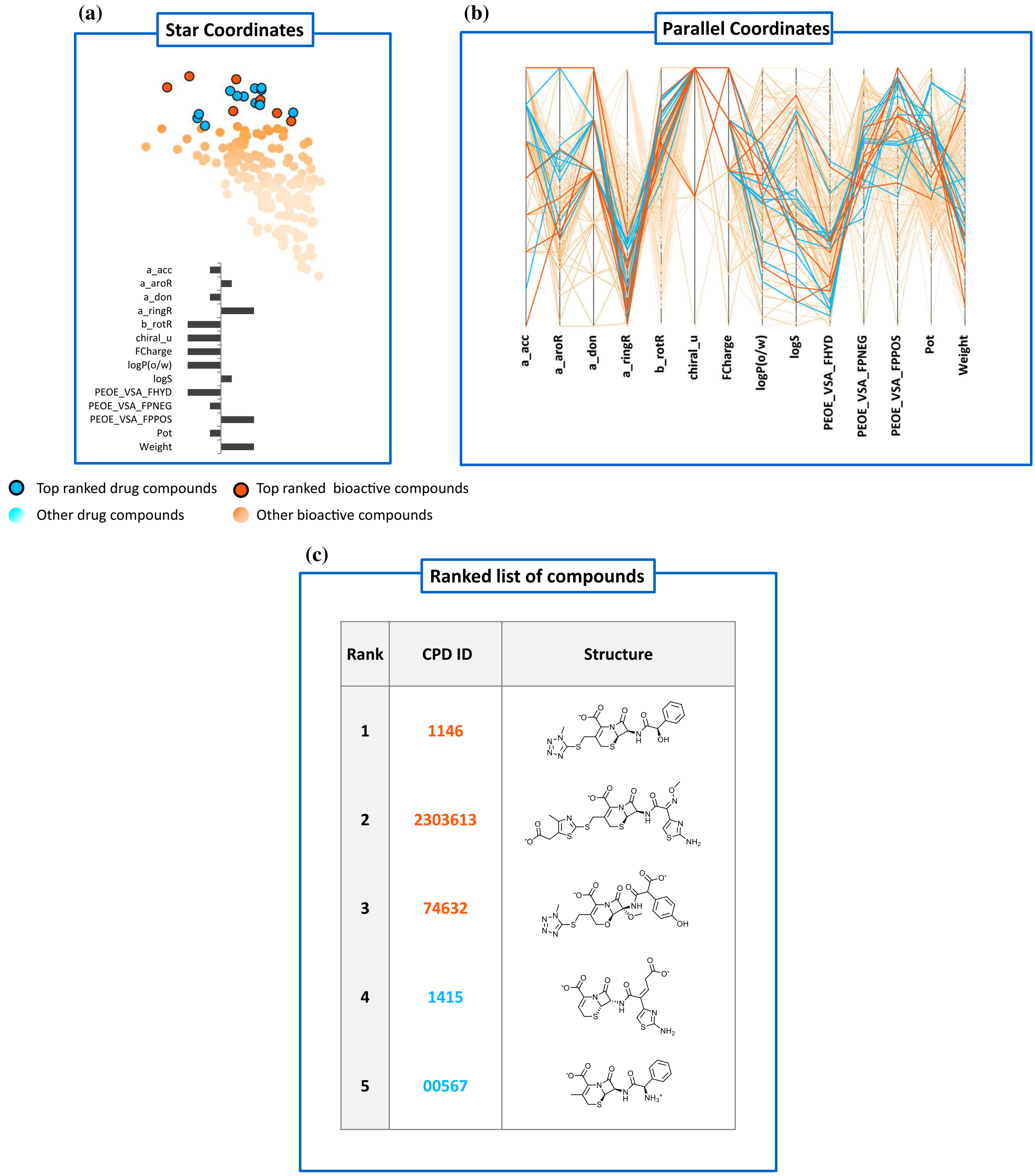


Fig. 3 Views of a multi-property landscape. Compound distributions of small intestine oligopeptide transporter ligands (ChEMBL target ID 4605) for a given projection were displayed using **a** STC and **b** PAC representations. In the STC representation, points represent individual compounds and color-coding distinguishes drugs (*cyan*) and bioactive compounds (*orange*). Top ranked molecules are depicted with a *black* border. In addition, *shading* of compounds indicates their rank, from *dark colors* (high rank, beginning at rank 21) to *light colors* (low rank). In the PAC representation, descriptors are assigned to vertical evenly *spaced lines* (spanning their value ranges) and compounds are depicted as *lines* (*horizontal traces*) *color-coded* as in (**a**). **c** The top five compounds from the ranking of the projection including two drugs (*4* and *5*). *Orange* and *cyan* compound (CPD) IDs correspond to ChEMBL and DrugBank IDs, respectively

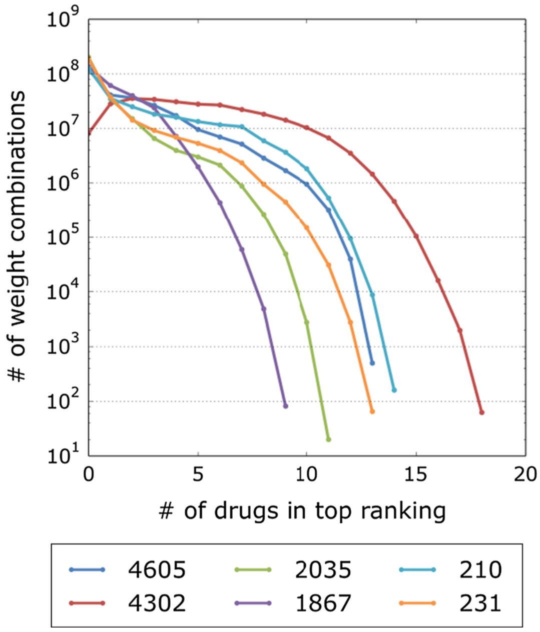


Fig. 4 Numerical comparison of projections. A projection was created for each weight value setting of the multi-objective function containing 14 descriptors and the number of drugs within the 20 top ranked compounds was determined. The *graph* reveals the number of weight combinations yielding largest numbers of highly-ranked drugs across the different target sets (*colored* by target IDs given in Table 1)

spread the compound data set across the property space, providing a clear view of compounds in increasingly large distances from the prioritized subspace. By contrast, in projection 2, the bulk of the data set was concentrated in a small region of property space and the separation of highly ranked and other compounds was only marginal. Hence, the property settings of projection 2 rendered data set compounds much more similar in multi-dimensional property space than the settings of projection 1, as clearly revealed by STC visualization. Therefore, for the selection of candidate compounds for chemical optimization efforts focusing on drug-like subspace, preference would be given to projection 1.

In Fig. 5c, two projections with similar descriptor weights are shown for alpha-2a adrenergic receptor ligands that yielded large drug enrichment (and shared seven of eight drugs among the top 20 compounds). Figure 5d shows the STC visualizations of these projections. Although the weight combinations were very similar, the compound distributions were distinct, contrary to expectations, as further discussed below.

For comparison with STC, Fig. 5e shows PC plots (using the first and second PC) of the unweighted descriptor space and weighted descriptor combinations of projections 1 and 2. In unweighted descriptor space, PCA did not yield a separation of drugs and bioactive compounds. Moreover, the PC plots of projection 1 and 2 were very difficult to interpret and remained essentially inconclusive. By contrast, the STC representations of projection 1 and 2 in Fig. 5d reveal a clear separation of top ranked and other data set compounds, but with different characteristics. The STC view of projection 1 shows that many data set compounds including remaining drugs were located proximal to the prioritized subspace, while only a small number of lowly ranked compounds were far removed from it. However, the STC view of projection 2 in Fig. 5d reveals a significant spread of the compounds across multi-dimensional property space (similar to projection 1 in Fig. 5b) including the majority of drugs, although the weight settings of projection 1 and 2 were comparable. In the case of projection 2, the STC view also shows that the drug-like subspace was less well-defined than in other cases, with many drugs (including two highly ranked ones) located distantly from many top ranked compounds. From these STC views, individual compounds can be easily selected for further analysis. Taken together, the STC visualizations provided a well-resolved picture of compound distributions in multi-dimensional property space for otherwise very similar projections.

Concluding remarks

High-dimensional property spaces for compound optimization or data set analysis are generally difficult to represent and navigate. While the potency-centric AL concept has substantially contributed to graphical SAR exploration, especially for larger and structurally heterogeneous data sets, little efforts have thus far been made to visualize multi-dimensional property landscapes that combine activity with other optimization-relevant properties. Typically, dimension reduction techniques such as PCA are applied to evaluate feature contributions in multi-dimensional space. Different types of graphical analysis are expected to aid in the rationalization of multi-dimensional property spaces. Therefore, a visualization methodology for multi-dimensional property spaces has been developed, as reported herein. Our analysis was based upon the generation of drug-like subspaces in chemical space, which takes molecular similarity relationships implicitly into account. However, it would also be feasible to focus an analysis explicitly on selected distance relationships in chemical space (or generate subspaces for compound reference sets with other characteristic properties).

Our study introduces the STC and PAC concepts, adapted from computer graphics, to the medicinal chemistry community. STC/PAC visualization of compound data is designed to complement multi-objective optimization, provide access to multi-dimensional data distributions, and aid in compound selection. For a given

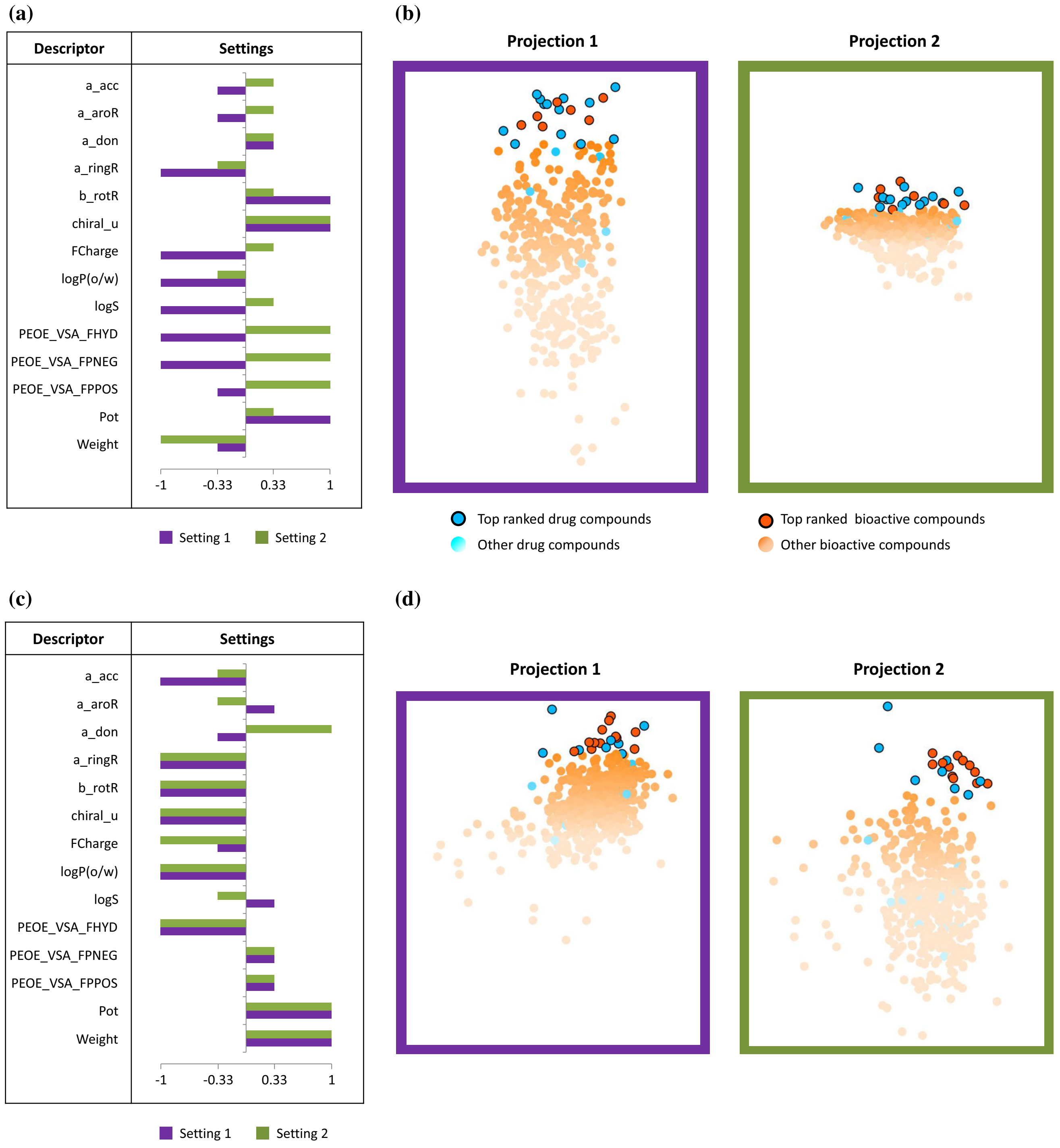


Fig. 5 Visualization of projections. Exemplary projections are visualized and compared. In (**a**) and (**b**), two projections generated for beta-2 adrenergsic receptors (ChEMBL target ID 210) are shown. The corresponding top 20 rankings contained 13 drugs each (11 of which were the same). **a** Compares the weight combinations (settings) for these projections and **b** their STC visualizations. *Points* represent individual compounds and are *color-coded* according to Fig. 3a. In (**c**) and (**d**), two projections generated for alpha-2a adrenergic receptor ligands (ID 1867) are shown. The corresponding top 20 rankings contained eight drugs each (seven of which were the same). **c** Compares the weight combinations (settings) for these projections and **d** their STC visualizations. In (**b**) and (**d**), STC visualizations were scaled to the same value ranges. **e** PCA-based data set projections (using the first two PCs) with unweighted descriptors (*top*, drugs colored *cyan* and bioactive compounds *gray*) and weighted descriptors from projection *1* (*middle*) and 2 (*bottom*) taken from (**c**). PCA plots of projections are color-coded as in (**d**)

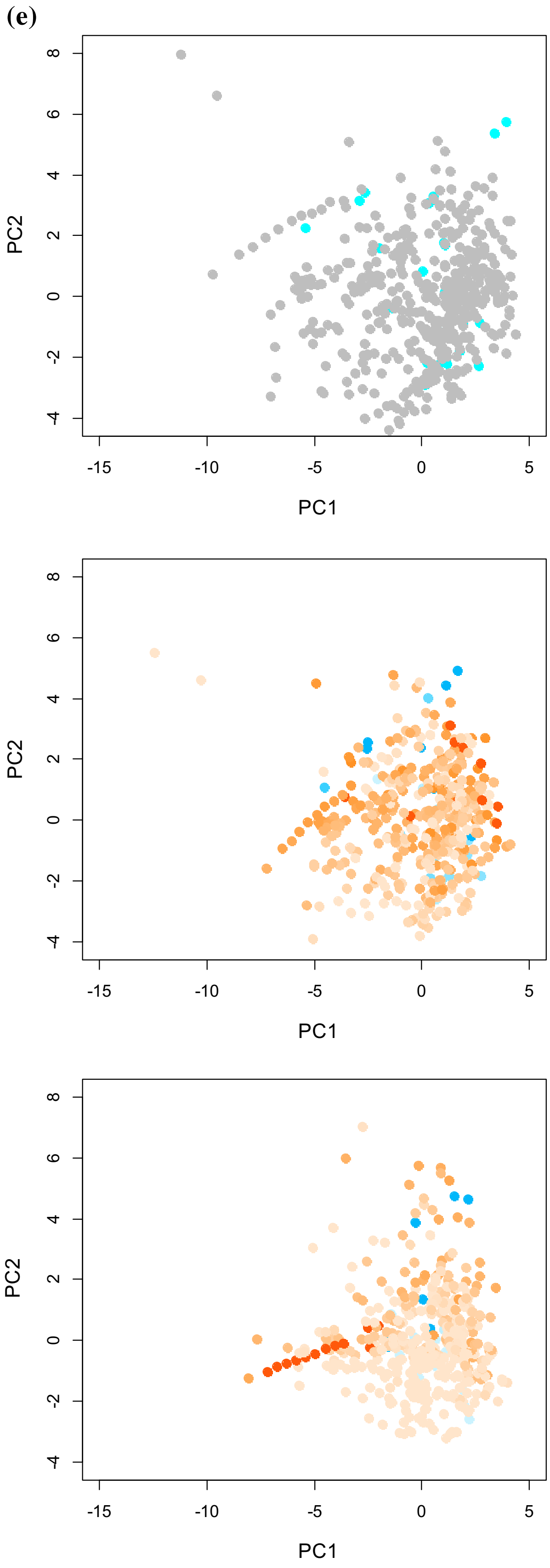


Fig. 5 continued

projection and compound ranking, the STC visualization provides a 2D representation of a compound distribution in multi-dimensional property space and views highly ranked compound subsets in the data set context. In addition, the PAC representation compares individual property contributions and identifies property settings that distinguish highly ranked compounds from others. We have demonstrated that STC visualizations help to differentiate numerically equivalent optimization solutions with similar or distinct property settings. The data sets used herein are made freely available [30].

References

1. Stumpfe D, Bajorath J (2012) Methods for SAR visualization. RSC Adv 2:369–378
2. Wassermann AM, Wawer M, Bajorath J (2010) Activity landscape representations for structure-activity relationship analysis. J Med Chem 53:8209–8223
3. Shanmugasundaram V, Maggiora GM (2001) Characterizing property and activity landscapes using an information-theoretic approach. In: Proceedings of 222nd American chemical society national meeting, division of chemical information, Chicago, IL, August 26–30, 2001; American Chemical Society: Washington, D.C., 2001; abstract no. 77
4. Wawer M, Peltason L, Weskamp N, Teckentrup A, Bajorath J (2008) Structure-activity relationship anatomy by network-like similarity graphs and local structure-activity relationship indices. J Med Chem 51:6075–6084
5. Wollenhaupt S, Baumann K (2014) inSARa: Intuitive and interactive SAR interpretation by reduced graphs and hierarchical MCS-based network navigation. J Chem Inf Model 54:1395–1409
6. Agrafiotis DK, Shemanarev M, Connolly PJ, Farnum M, Lobanov VS (2007) SAR maps: a new SAR visualization technique for medicinal chemists. J Med Chem 50:5926–5937
7. Wassermann AM, Bajorath J (2012) Directed R-group combination graph: a methodology to uncover structure-activity relationship patterns in a series of analogues. J Med Chem 55:1215–1226
8. Peltason L, Weskamp N, Teckentrup A, Bajorath J (2009) Exploration of structure-activity relationship determinants in analogue series. J Med Chem 52:3212–3224
9. Wawer M, Bajorath J (2010) Similarity-potency trees: a method to search for SAR information in compound data sets and derive SAR rules. J Chem Inf Model 50:1395–1409
10. Peltason L, Iyer P, Bajorath J (2010) Rationalizing three-dimensional activity landscapes and the influence of molecular representations on landscape topology and the formation of activity cliffs. J Chem Inf Model 50:1021–1033
11. Reutlinger M, Guba W, Martin RE, Alanine AI, Hoffmann T, Klenner A, Hiss JA, Schneider P, Schneider G (2011) Neighborhood-preserving visualization of adaptive structure-activity landscapes: application to drug discovery. Angew Chem Int Ed 50:11633–11636
12. Zwierzyna M, Vogt M, Maggiora GM, Bajorath J (2015) Design and characterization of chemical space networks for different compound data sets. J Comput-Aided Mol Des 29:113–125
13. Ertl P, Rohde B (2012) The molecule cloud-compact visualization of large collections of molecules. J Cheminf 4:12

14. Awale M, van Deursen R, Reymond J-L (2010) MQN-mapplet: visualization of chemical space with interactive maps of DrugBank, ChEMBL, PubChem, GDB-11, and GDB-13. J Chem Inf Model 50:1395–1409
15. Reymond J-L (2015) The chemical space project. Acc Chem Res 48:722–730
16. Kireeva N, Baskin II, Gaspar HA, Horvath D, Marcou G, Varnek A (2012) Generative topographic mapping (GTM): universal tool for data visualization, structure-activity modeling, and dataset comparison. Mol Inf 3(4):301–312
17. Wermuth CG (2008) The practice of medicinal chemistry, 3rd edn. Academic Press-Elsevier, Burlington, London
18. Gillet VJ, Khatib W, Willett P, Fleming P, Green DVS (2002) Combinatorial library design using multiobjective genetic algorithm. J Chem Inf Comput Sci 42:375–385
19. Gillet VJ (2004) Applications of evolutionary computation in drug design. Struct Bond 110:133–152
20. Nicolaou CA, Brown N, Pattichis CS (2007) Molecular optimization using computational multi-objective methods. Curr Opin Drug Discov Develop 10:316–324
21. Gaulton A, Bellis LJ, Bento AP, Chambers J, Davies M, Hersey A, Light Y, McGlinchey S, Michalovich D, Al-Lazikani B, Overington JP (2012) ChEMBL: a large-scale bioactivity database for drug discovery. Nucleic Acids Res 40:D1100–D1107
22. Law V, Knox C, Djoumbou Y, Jewison T, Guo AC, Liu Y, Maciejewski A, Arndt D, Wilson M, Neveu V, Tang A, Gabriel G, Ly C, Adamjee S, Dame ZT, Han B, Zhou Y, Wishart DS (2014) DrugBank 4.0: shedding new light on drug metabolism. Nucleic Acids Res 42:D1091–D1097
23. OEChem TK (2012) OpenEye scientific software Inc, Santa Fe, NM, USA
24. Molecular Operating Environment (2012) Chemical computing group Inc.: Montreal, Quebec, Canada
25. Cook D, Buja A, Lee EK, Wickham H (2008) Grand tours, projection pursuit guided tours and manual controls. In: Chen C, Härdle W, Unwin A (eds) Handbook of data visualization. Springer, Heidelberg, pp 295–314
26. Kandogan E (2000) Star coordinates: a multi-dimensional visualization technique with uniform treatment of dimensions. In: LBHT Proc IEEE information visualization symposium, pp 9–12
27. Java universal network/graph framework. http://jung.sourcefourge.net/. Accessed May 1, 2014
28. Inselberg A (1985) The plane with parallel coordinates. Visual Comput 1:69–91
29. R: a language and environment for statistical computing. R foundation for statistical computing, Vienna, Austria, 2012
30. de la Vega de León A, Kayastha S, Dimova D, Schultz T, Bajorath J (2015) ChEMBL20 data sets for multi-property landscape analysis. ZENODO. doi:10.5281/zenodo.21782

Summary

We have presented a proof-of-concept study demonstrating the application of star coordinates to visualize projections arising from multi-property optimization and parallel coordinates to compare relationships among underlying descriptors. A 14-dimensional space comprising a set of chemically intuitive descriptors including compound potency was designed as a reference space for exemplary multi-property optimization. Bioactive compounds and drugs were projected from multi-dimensional to one-dimensional space using multi-objective function. A systematic search was performed by using four different weights for each descriptor to find the combination that maximizes the the function value, based on which compounds were ranked. Projections with maximum number of drugs on top were further prioritized. The multi-objective function hence corresponded to drug-likeliness of compounds. Star coordinates was used to visualize individual projections. Although comparable drug enrichment was obtained for distinct descriptor weight combinations, star coordinate visualization revealed significant difference in their compound distribution in multi-property landscape. Further, projections with similar descriptor weight settings and sharing most of the top ranked drugs also had distinct compound distribution in multi-dimensional property space. Descriptor correlation were analyzed in parallel coordinate plots, whereby descriptor values of drugs and bioactive compounds could be compared.

I have contributed to the implementation of STC visualization tool, mainly for generating the ranked list of compounds based on their MOF values.

Computational multi-property optimization methods are mostly employed in compound library design or in search for drug-like subspaces in chemical space. However, in practical lead optimization scenario, medicinal chemists mainly focus on deciding the next compound to be synthesized. The ultimate success or failure of a given project depends on the compounds comprising the lead series. In the following chapter, we present the application of SAR matrix data structure to quantify SAR progression during lead optimization with an aim to guide medicinal chemists for better decision making.

Chapter 8

Monitoring the Progression of Structure-Activity Relationship Information during Lead Optimization

Introduction

Lead optimization involves synthetic exploration of selected active compounds comprising lead series to improve their drug-relevant properties. It is during this stage of drug discovery that prioritized compounds are transformed into pre-clinical drug candidates via iterative analog evaluation. Compound optimization, as we have discussed already is highly challenging, taking into consideration its multiparameteric nature. Apart from improving potency of compounds, a balance needs to be reached among multiple properties, that would ultimately render a given compound drug-like. Although computational multi-property optimization methods have been developed, they are mostly popular in guiding compound library design or reducing the multi-property space to focus on drug-like subspaces. Medicinal chemists involved in lead opmimization projects are typically less interested in solutions proposed by multi-property optimization approaches. This might be partly due to the fact that such solutions cannot be easily interpreted in terms of chemical reasoning and is rather vague from medicinal chemistry perspective.

The decisions made in the course of refining a given set of analogs are largely based on the experience and chemical intuition of the medicinal chemists leading the lead optimization projects. In addition, lead optimization projects generally last for several years, during which hundreds and thousands of compounds are synthesized and immense amount of resources are expended. It cannot be easily predicted if a given lead series will ultimately give rise to a successful clinical candidate or not. In many cases, projects might even need to be terminated. It is therefore highly desirable to have a method that can objectively assess the progression of a lead optimization project. Nevertheless, only few computational approaches have been developed to these ends. Among them are the similarity network-based evaluation of SAR progression, in which compound communities comprising SAR rich and flat SAR characteristics could be spotted. More recently, a statistical framework to calculate risk associated with a given compound series was proposed.

In this study, we apply SAR matrices (SARMs) as a diagnostic tool to evaluate the progression of SAR information in lead optimization data sets. SARMs generated at different time points of evolving data sets, termed here as indicator SARMs, capture structural relationships among analogs. Median potency and SAR discontinuity scores calculated for individual indicator SARMs were used as metrics to evaluate positive, negative or neutral SAR progression.

Journal of Medicinal Chemistry

Article

pubs.acs.org/jmc

Monitoring the Progression of Structure−Activity Relationship Information during Lead Optimization

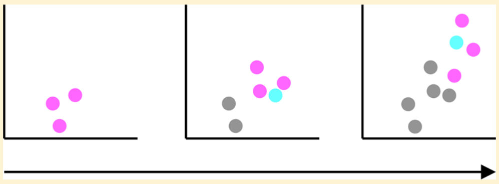
Veerabahu Shanmugasundaram,[†] Liying Zhang,[‡] Shilva Kayastha,[§,∥] Antonio de la Vega de León,[§,∥] Dilyana Dimova,[§] and Jürgen Bajorath*[,§]

[†]Center of Chemistry Innovation & Excellence, WorldWide Medicinal Chemistry, Pfizer PharmaTherapeutics Research & Development, Eastern Point Road, Groton, Connecticut 06340, United States

[‡]Computational Sciences CoE, WorldWide Medicinal Chemistry, Pfizer PharmaTherapeutics Research & Development, 610 Main Street, Cambridge, Massachusetts 06340, United States

[§]Department of Life Science Informatics, B-IT, LIMES Program Unit Chemical Biology and Medicinal Chemistry, Dahlmannstr. 2, Rheinische Friedrich-Wilhelms-Universität, D-53113 Bonn, Germany

ABSTRACT: Lead optimization (LO) in medicinal chemistry is largely driven by hypotheses and depends on the ingenuity, experience, and intuition of medicinal chemists, focusing on the key question of which compound should be made next. It is essentially impossible to predict whether an LO project might ultimately be successful, and it is also very difficult to estimate when a sufficient number of compounds has been evaluated to judge the odds of a project. Given the subjective nature of LO decisions and the inherent optimism of project teams, very few attempts have been made to systematically evaluate project progression. Herein, we introduce a computational framework to follow the evolution of structure−activity relationship (SAR) information over a time course. The approach is based on the use of SAR matrix data structures as a diagnostic tool and enables graphical analysis of SAR redundancy and project progression. This framework should help the process of making decisions in close-in analogue work.



INTRODUCTION

Lead optimization (LO) aims to transform selected active compounds into clinical candidates through iterative close-in analogue evaluation and is one of the most important challenges in the practice of medicinal chemistry.[1] To date, the multiparametric LO process[1] has been largely driven by a combination of hypotheses and empirical rules that vary based on chemical intuition and experience. The key question faced by medicinal chemists during LO is which compound(s) should be made next, and educated guesses about suitable chemical modifications typically provide the basis for generating analogues and advancing LO projects.

In addition to improving compound potency and selectivity, other properties that are also considered during optimization include solubility, permeability, metabolic stability, and bioavailability. Balancing multiple compound properties in the course of lead optimization is a significant challenge that strongly depends on the specifics of the therapeutic applications and compound classes under study.

Given the multiparametric nature of LO, computational approaches focusing on multiobjective optimization have been developed to aid compound design.[2,3] These methods often employ desirability functions or probability estimates to model and balance multiple drug-relevant properties and select computationally designed candidate compounds with preferred property profiles.[3] However, it is probably fair to say that advanced multiobjective optimization is more popular in library design efforts or in limiting an area of property space on which to focus rather than practical LO, where the pivotal *which compound should be made next* question rules day-to-day efforts.

LO projects often require long periods of time and a large amount of resources. It is not uncommon for hundreds or thousands of compounds to be generated over the course of several years by project teams pursuing multiple lead series, often while facing many roadblocks along the way. In light of this situation, it is difficult to objectively assess LO progression. If a project faces roadblocks, then there is always hope that the next compound(s) might present a breakthrough. This optimism might carry a LO project for a long period of time, and the more time and effort that are expended on it, the more difficult it typically becomes to let go and terminate a project due to limited success. It is therefore not surprising that medicinal chemistry leaders are equally concerned about positive, neutral, or negative project progression and that questions such as how many more compounds do we need to make in close-in analogue space until we reach a go/no-go decision are common place in industry. Accordingly, metrics to assess and quantify LO project progression in a more objective manner are highly desirable. However, only small advances have thus far been made to conceptualize and implement such metrics for the practice of medicinal chemistry.

Special Issue: Computational Methods for Medicinal Chemistry

Received: September 15, 2015

DOI: 10.1021/acs.jmedchem.5b01428

Although many computational methods for compound design and activity prediction are available, only very few attempts have been reported to computationally evaluate LO progression, a task that principally differs from compound design. For example, structure−activity relationships (SARs) contained in evolving compound data sets have been monitored in molecular network representations annotated with activity information as well as using three-dimensional activity landscape models.[4] In similarity-based compound networks, positive SAR progression over time is reflected by the formation of compound communities rich in SAR information, whereas lack of progression is indicated by increasing numbers of compounds populating flat SAR regions.[4] Comparison of networks generated at different time points of a project provides a qualitative view of SAR progression. However, the interpretation of SAR networks is not trivial for non-experts.

Furthermore, in a recent investigation, a statistical framework for assessing LO progress has been introduced.[5] For multiple LO parameters, the risk associated with a compound set is quantified from value distributions as the deviation from desired threshold values, and the global risk is obtained by combining all parameter contributions. During the LO process, the risk is expected to be minimized. Risk as a function of (temporal) project progression can be graphically analyzed in different ways, and key compounds making the largest contributions to risk minimization can be identified.[5] Pros of this statistical approach include the ability to monitor multiple properties, individually or in concert, and that it quantifies risk; cons include the requirement of the approach to define property thresholds and that it does not take structural information or relationships as parameters into account (for similarity or diversity assessment, additional computational methods must be employed). Therefore, it is not designed for systematic SAR exploration. In another recent investigation, LO attrition analysis has been introduced[6] to classify compounds according to the number of LO criteria they meet. For this purpose, (project-specific) preferred ranges of numerical properties must be defined and expressed as binary yes/no queries, and the number of compounds meeting an increasing number of queries is determined. Attrition curves are generated by plotting compound count vs parameter count (i.e., x compounds meet y parameters) and used to evaluate LO success.[6] As presented, the approach does not include a temporal component to monitor progress. For a given LO set, the attrition curves are suitable to provide a global view of compound quality. Further analyses performed thus far do not capture the totality of SAR information content for available analogues but, rather, debate the merits of each compound individually.

In this study, we introduce a conceptually different method for the evaluation of SAR progression during LO. The SAR matrix (SARM) data structure[7,8] originally developed for elucidation of SAR patterns in analogue series[7] has been adapted as an indicator of SAR information content for temporal analysis of LO data sets. SARM ensembles are calculated for evolving data sets and scored to quantify their SAR information content. In addition, matrices are classified according to the structural information they capture, which makes it possible to monitor the expansion of existing compound series as well as the introduction of structural novelty during LO in close-in analogue space. SARM distributions are graphically analyzed, and changes in distributions over time reveal SAR progression or a lack of progression. Indicator SARMs can also be annotated with multiple properties, and changes in property profiles can be monitored. Since SARMs exhaustively dissect compound sets in a systematic manner, it is envisioned that the wealth of SAR information during LO might be revealed through an analysis of SARM ensembles over a time course.

■ EXPERIMENTAL SECTION

SARM Generation. SARMs are generated after subjecting compound sets to two-stage matched molecular pair (MMP) generation.[7,8] A MMP is defined as a pair of compounds that differ only by a structural modification at a single site.[9] MMPs are efficiently generated by systematic fragmentation of exocyclic single bonds in compounds (permitting single, double, and triple cuts) and collection of core structures and associated substituents in index tables.[10]

In the first step, MMPs are generated for all compounds. In the second step, which is uniquely applied for SARMs, all core structures resulting from the first round of fragmentation are again subjected to MMP generation. Compounds forming MMPs from the first step are organized as matching molecular series (MMSs). A MMS is defined as a series of compounds that share the same core and have different substituents at a single site (representing an extension of the MMP concept).[11] It follows that compounds comprising an MMS must form all possible pairwise MMPs. Each MMS is represented as the shared core plus the set of distinguishing substituents. Core MMPs from the second round of fragmentation then identify all structurally analogous cores (differing only by a change at a single site). Each SARM contains a unique subset of MMSs with structurally analogous cores. In the matrix, each row represents an MMS with a unique core (and each column represents a substituent). As a consequence of systematic MMP fragmentation, compounds typically participate in multiple MMSs and occur in multiple SARMs. The ensemble of SARMs generated from a compound set captures all possible analogue relationships. As shown in Figure 1A, SARMs are reminiscent of conventional R-group tables. Each cell represents a unique combination of a core and substituent resulting from the fragmentation (including virtual compounds that have not yet been generated). Cells can be annotated with property information, for example, they can be color-coded according to compound potency, as also illustrated in Figure 1A.

Following the protocol outlined above, SARMs were generated with a Java program utilizing the OEChem toolkit.[12]

SAR Evaluation. The SAR information contained in a SARM was quantified by calculating two different values: the median potency of all compounds comprising the SARM and a matrix-based SAR discontinuity score (Figure 1B). SAR discontinuity is high when structurally similar or analogous compounds have significant potency variations.[13] Such compounds typically reveal SAR information. A SAR discontinuity score quantifying this information was first introduced by systematically accounting for pairwise potency differences between compounds meeting a predefined similarity criterion.[13] For SAR monitoring, we defined a SARM-based discontinuity score (SARM_Disc)

$$\text{SARM\_Disc} = \frac{\sum_{i}^{m} \sum_{j,j>i}^{m} |\text{pot}_i - \text{pot}_j|}{N} \; \forall\ i, j \rightarrow \text{MMP}$$

where i and j are compounds in a SARM that form an MMP, m is the total number of SARM compounds, N is the total number of MMPs contained in the SARM, pot_i is the potency of compound i, and pot_j is the potency of compound j. For each SARM, the SARM_Disc value was calculated.

Graphical Analysis. SARM distributions were analyzed in scatterplots of median potency vs SARM_Disc scores. In addition, trend plots were generated from SARM distributions to separately monitor the progression of potency and SARM_Disc scores over time. Trend plots were obtained by fitting potency and SARM_Disc values averaged at different time intervals to a linear function.

DOI: 10.1021/acs.jmedchem.5b01428

Table 1. ChEMBL Compound Data Sets and SAR Matrices[a]

			first year		last year	
ID	target name	years	no. cpds	no. SARMs	no. cpds	no. SARMs
1908	cytochrome P450 11B1	2006−2013	68	7	464	206
4015	C−C chemokine receptor type 2	2006−2011	124	182	836	1365
344	melanin concentrating hormone receptor 1	2005−2010	259	329	990	1086
3468	caspase-7	2005−2014	61	13	232	125

[a] For each data set, the ChEMBL ID and target name are reported as well as the time period (years) over which the growth of the data set was monitored using SARM ensembles. In addition, the compound composition (no. cpds) and corresponding SARM statistics (no. SARMs) are provided for the first and last years of each time period.

Table 2. Pfizer LO Data Sets and SAR Matrices[a]

			first year		last year	
LO targets and sets		years	no. cpds	no. SARMs	no. cpds	no. SARMs
neurodegenerative	series 1	2010−2014	10	1	431	672
	series 2	2010−2015	46	49	125	128
inflammation	series 1	2011Q1−2012Q3	20	5	88	93
	series 2	2010Q2−2010Q4	18	9	78	43

[a] For each LO set, the time period (years) is reported over which the growth of the corresponding compound series was monitored using SARM ensembles. Q means quarter. In addition, the compound composition (no. cpds) and corresponding SARM statistics (no. SARMs) are provided for the first and last intervals of each time period.

Public Domain Data Sets. Compounds and activity data were taken from ChEMBL[14] (version 20). To assemble data sets evolving over time, compounds for proof-of-concept studies active against human targets at the highest confidence level (ChEMBL confidence score 9) with reported direct binding interactions (ChEMBL relationship type D) and IC_{50} values as potency measurements were considered. For all preselected compounds, publication dates were recorded. A qualifying target-based data set was required to contain compounds reported in increments over a period of at least 5 subsequent years (for each year, a new compound subset had to be available), with a minimum of 50 compounds available in the first year. Four data sets meeting these criteria were assembled, as reported in Table 1.

LO Data Sets. In addition to ChEMBL sets, two LO data sets originating from two different drug discovery projects at Pfizer were studied. Each project team pursued two different chemical series. In each case, one of the series was deemed to be a successful chemical series because the project team was able to identify and nominate preclinical candidate(s), and the second was an unsuccessful series from which no candidate compound was nominated. The first target protein was an enzyme, which was pursued as a biological target for a neurodegenerative indication. The end point for potency in this project was inhibitory activity assessed in a direct enzymatic assay. Although the project team also evaluated other properties during LO, for the purposes of this study, the primary potency end point was used to monitor SAR progression. The second target was also an enzyme, and downregulation of the activity of this enzyme was targeted for an inflammation indication. Also in this case, the end point for potency was inhibitory activity in an enzymatic assay. A series definition used by the project team was added to each compound. IC_{50} values for both projects were converted to logarithmic units. For temporal analysis, dates when compounds were first registered internally were determined and used for monitoring SAR progression. Details of the LO data sets are reported in Table 2.

RESULTS AND DISCUSSION

Concept of Indicator SARMs. SARMs were originally developed for a completely different purpose than for monitoring SAR progression during LO, i.e., to systematically organize analogue series, elucidate SAR patterns for structurally related series, suggest virtual compounds, and predict their activity.[8] In Figure 1A, a small model SARM formed by six compounds (two MMSs) is shown on the left, and a slightly larger SARM (seven compounds, two MMSs) is shown on the right, which also contains a virtual compound (non-colored cell). We reasoned that several characteristics of SARMs might render them suitable for monitoring SAR progression:

(1) SARMs systematically extract all analogue relationships from compound sets. If LO sets contain multiple series, then SARMs not only organize these series as MMSs but also detect all structural relationships among them. Each SARM contains a unique subset of MMSs with related core structures, regardless of the origin of these structural relationships.

(2) SARMs can be easily annotated with compound properties that can then be analyzed based upon the structural organization provided by SARMs.

(3) Depending on the structural relationships contained in a compound data set, varying numbers of SARMs are obtained. This is illustrated in Table 1, which reports compound and SARM statistics for the public domain data sets. Since LO sets are typically centered on single or multiple lead series, they tend to produce large SARM ensembles, thus enabling statistical analysis of SARMs and SARM-associated properties. As a rule-of-thumb, the number of SARMs obtained for structurally homogeneous data sets is often roughly comparable to the number of data set compounds (Table 1).

Given these characteristics, we introduced three modifications to SARMs specifically for the purpose of SAR progression analysis:

(1) SARMs were iteratively calculated for evolving compound data sets at different time points. Thereby, SARM ensembles were obtained that systematically captured all structural relationships between existing and new compounds.

(2) For the analysis of these ensembles, SARMs were classified into three categories including *existing*, *expanded*, and *new* SARMs. Existing SARMs were not modified through the addition of new compounds,

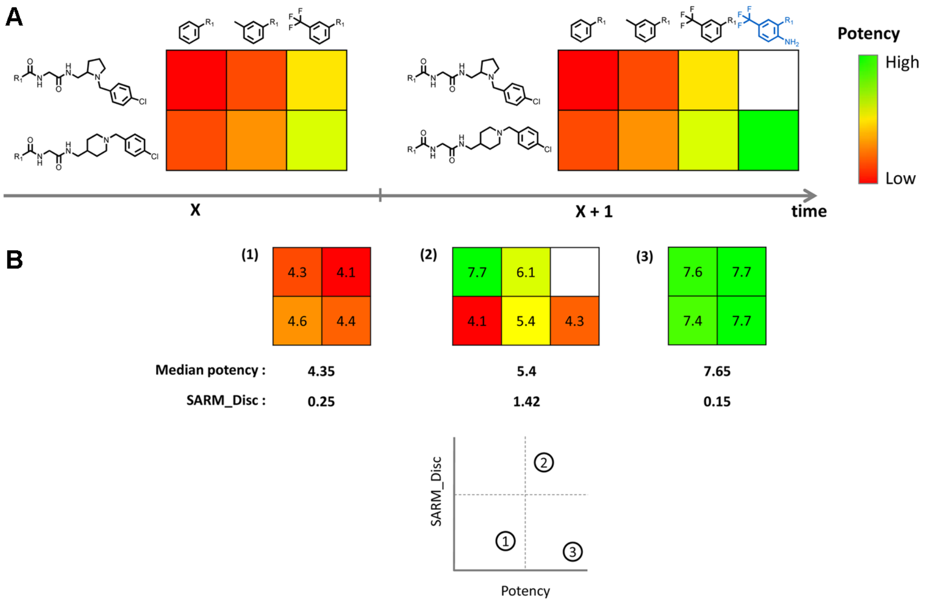


Figure 1. SARM, expansion, and characterization. (A) In the SARM, each row represents a matching molecular series (MMS), i.e., a series of compounds that have a common core (shown left from the row) and are distinguished only by a substituent at a single site (top of each column). Each cell represents an individual compound (unique combination of a core and substituent), either a known data set compound (colored by potency using a continuous spectrum from (lowest) red to (highest) green) or a virtual compound (an as of yet unexplored combination of a core and substituent; non-colored cell). All MMSs contained in a given SARM have related cores that are distinguished only by a structural change at a single site. The matrix on the left was expanded through the addition of a new compound that was detected to match the core of one of the MMSs contained in this matrix. The resulting expanded matrix is shown on the right (the substituent of the new compound is highlighted in blue). (B) Exemplary SARMs with varying SAR information content. SARMs were characterized by calculating their median compound potency and the SARM_Disc score (see text). Accordingly, the SARM_Disc score of a SARM is high if the structurally related compounds comprising the SARM have large potency variations. Therefore, SARM_Disc scores serve as an indicator of SAR information content. As can be seen (and easily rationalized), median potency does not per se correlate with SARM_Disc. The three exemplary SARMs are shown in a scatterplot of median potency vs SARM_Disc. The scatterplot is divided into four quadrants. SARMs with high information, such as matrix 2 in this example, map to the upper right quadrant.

whereas expanded SARMs were obtained when new compounds form structural relationships with already available compounds (as is the case when new analogues are generated for an existing series). Figure 1A illustrates the process of SARM expansion. A new compound complements one of the two MMSs contained in the matrix on the left, leading to the generation of an expanded SARM on the right. Moreover, if newly added compounds introduced structural novelty, i.e., if they formed novel MMSs, then new SARMs were obtained.

(3) For SAR monitoring, SARMs were annotated with two properties, including compound potency and the newly introduced SARM-based SAR discontinuity score (SARM_Disc), as illustrated in Figure 1B. For each SARM, the median potency and the SARM_Disc score were calculated. A high SARM_Disc score indicated the presence of structural analogues with significant potency variations. This situation corresponded to high SAR information content of a SARM because it encoded structural changes that significantly affected potency (different from SARMs that exclusively consisted of weakly or highly potent analogues). Taken together, median potency and SARM_Disc made it possible to prioritize matrices for SAR monitoring. From a SAR information perspective, progress during LO is generally made when SAR-sensitive analogues are obtained including increasingly potent compounds during the course of the project. Following our analysis concept, this is reflected by the generation of SARMs with high median potency and high SARM_Disc scores (as an inflection point during the course of the project), as revealed by time-dependent analysis of matrix distributions.

SARM distributions were recorded in scatterplots of median potency vs SARM_Disc, as schematically represented in Figure 1B (bottom). Preferred SARMs with high median potency and high discontinuity scores mapped to the upper right quadrant of these plots.

The original SARM approach was focused on exploring individual matrices and the compound information that they contained, as discussed above. Because we did not consider the content of individual SARMs for monitoring SAR progression but studied SARM distributions with respect to property values over time, matrix ensembles generated for our current analysis were termed *indicator SARMs*.

Graphical SARM Distribution Analysis. Figure 2A summarizes the principles of time-dependent indicator SARM analysis. SARM ensembles were calculated for an evolving data set following each addition of a compound subset and classified according to the compounds and structural relationships that they captured. The resulting SARM distributions were monitored over time in scatterplots reflecting their SAR information content. Figure 2B shows exemplary progression trends. At the top, positive SAR progression is illustrated. In this case, matrix populations grew over time through the addition of new SARMs and, to a lesser extent, expanded

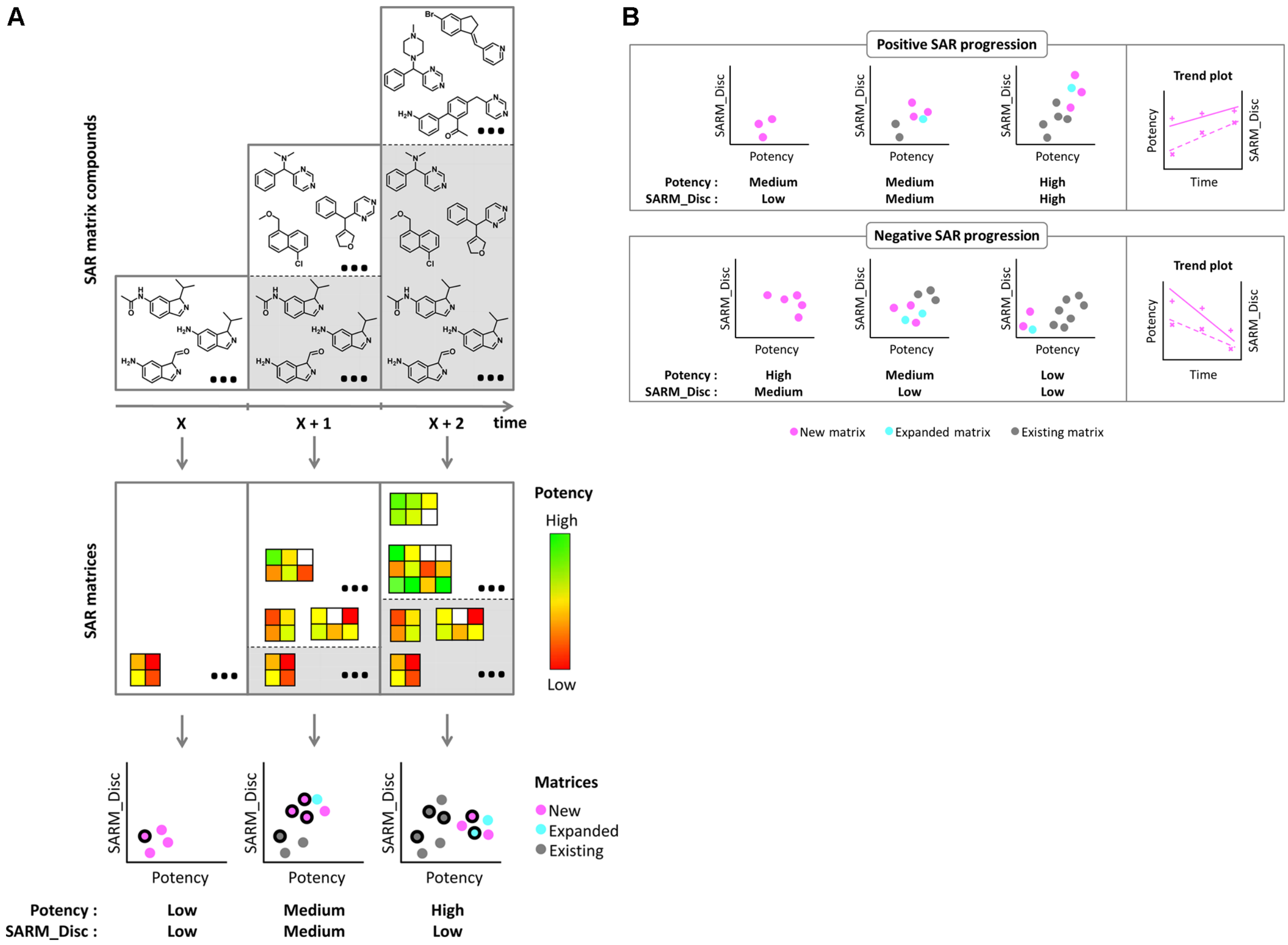


Figure 2. Monitoring SAR progression. (A) Schematic representation illustrating the concept of monitoring SAR progression over time using SARMs. Newly synthesized compounds (shown on a white background) are added in time intervals to evolving lead optimization sets (gray background), and SARMs are systematically calculated at each time point. Matrix representation is according to Figure 1. SARMs calculated at each time point are retained and compared to newly derived matrices. For visualization purposes, not all compounds and SARMs are shown. Distributions of SARMs are monitored in scatterplots of median potency vs SARM_Disc in which each SARM is represented as a color-coded dot. Dots with black border correspond to SARMs shown above the scatterplots. For temporal analysis, three categories of SARMs are distinguished: *existing* (colored gray), *expanded* (cyan), and *new* SARMs (magenta). Existing (old) matrices are not modified through the addition of newly synthesized compounds. Expanded SARMs evolve from existing matrices through the addition of analogues that further extend currently available MMSs. New SARMs contain new MMSs and capture previously unobserved structural relationships due to the addition of novel structures. (B) Two sets of SARM scatterplots are shown and color-coded as in panel (A). Comparison of SARM scatterplots makes it possible to follow SAR progression on a time course and judge the success of lead optimization (LO) efforts. For example, a desirable LO profile (top; positive SAR progression) would display a shift of matrix distributions over time toward the upper right quadrant of the scatterplot (characterized by the presence of high median potency and high SARM_Disc), with an enrichment of new SARMs. By contrast, the scatterplots at the bottom display negative progression of SAR over time because the matrix distribution shifts toward the bottom left quadrant (characterized by the presence of low median potency and low SARM_Disc). On the right, trend plots are shown obtained from indicator SARM distributions by fitting average potency and SARM_Disc scores of new matrices (magenta) for each year to linear functions. Trend lines monitor the development of SARM_Disc and potency for an indicator SARM category over time.

SARMs. A gradual shift of SARM distributions toward the upper right quadrants of the scatterplots was observed, revealing a steady increase in SAR information and the generation of increasingly potent compounds. By contrast, the example at the bottom illustrates (undesired) negative SAR progression characterized by the occurrence of expanded and new SARMs with low median potency and low discontinuity scores and the absence of an upward shift of SARM distributions over time. Positive and negative SAR progressions can also be visualized in trend plots (shown on the right of Figure 2B) that are derived from the SARM distributions by fitting linear models and separately monitoring potency and SARM_Disc progression over time. The trend lines were fitted to data averaged over time intervals. Ideally, in the case of positive SAR progression, these trend lines should have positive slopes.

Monitoring SAR Progression. Applying the approach summarized in Figure 2, SAR progression was monitored for different types of compound sets.

Public Domain Compound Sets. The four compound data sets from ChEMBL represented prototypic compound sets evolving over time and were generated to mimic LO sets by combining compounds active against different targets taken from the scientific literature (only high-confidence activity data

DOI: 10.1021/acs.jmedchem.5b01428

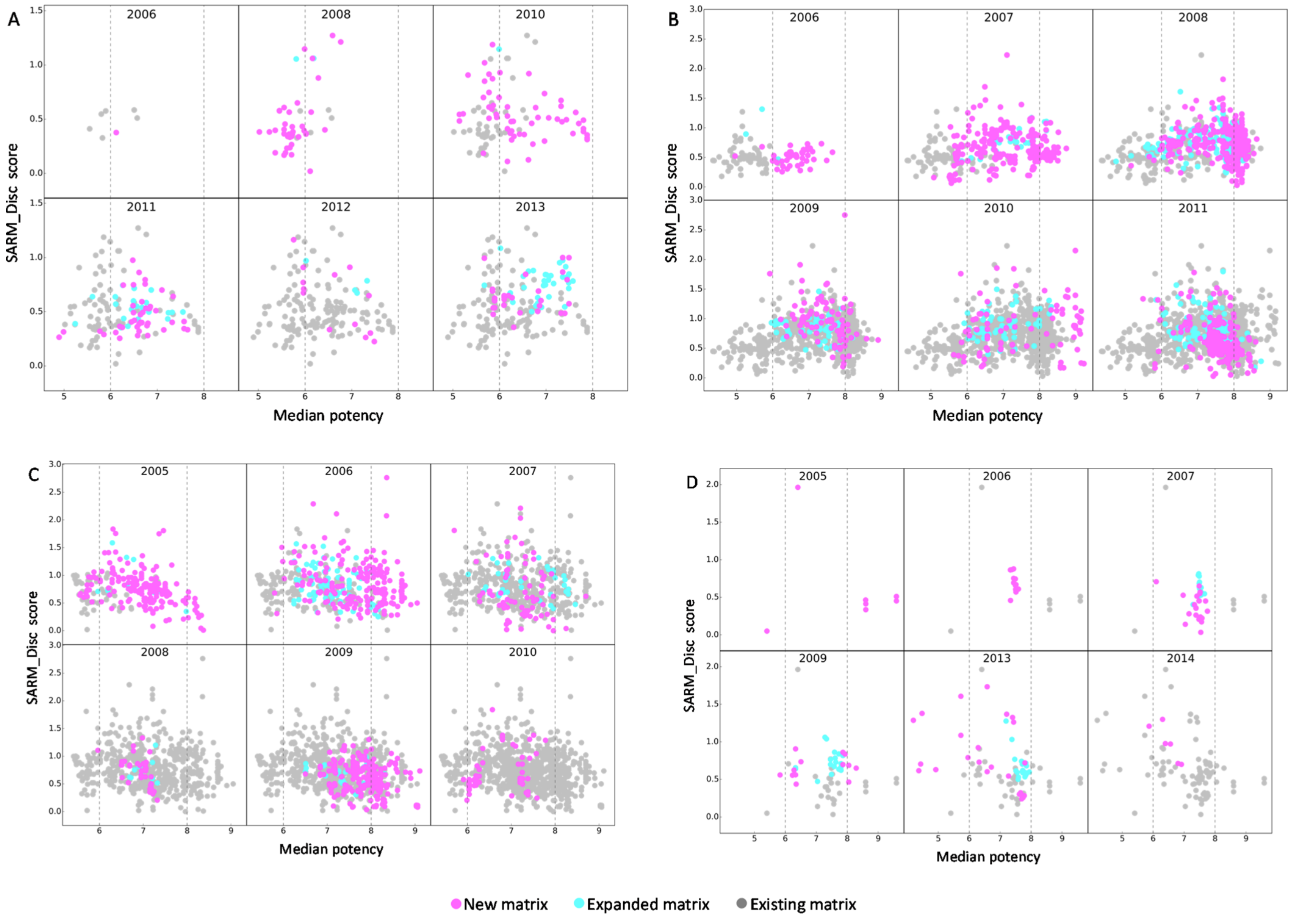


Figure 3. Indicator SARM distributions over a time course. Scatterplots are shown for four public domain data sets that were incrementally assembled over different years on the basis of compound publication dates. The SARM representation is according to Figure 2. In addition, dotted lines at potency values of six and eight log units differentiate SARMs with high, intermediate, or low median potency. (A) Cytochrome P450 11B1 inhibitors, (B) C−C chemokine receptor type 2 ligands, (C) melanin-concentrating hormone receptor 1 ligands, and (D) caspase-7 inhibitors. Compound and SARM statistics for the monitored time periods are provided in Table 1. We note that active compounds were available in each case prior to the first year monitored in a scatterplot. For compounds available in the preceding year, SARMs were calculated and used as a reference ensemble to generate classified SARMs for the first year of the monitored period.

were taken into consideration for compound selection). Because selected compounds originated from a variety of literature sources, these sets were structurally more heterogeneous than typical LO sets, thus presenting a challenge for a proof-of-concept assessment of indicator SARM analysis. These four data sets are made freely available as an open-access deposition.[15]

Figure 3 shows the distribution of indicator SARMs obtained from the data sets over a period of six subsequent years. The median potency and SARM_Disc scores of SARMs were plotted and colored according to their matrix category.

Figure 3A reports the temporal analysis of inhibitors of cytochrome P450 11B1. This set contained 464 compounds but yielded only 206 SARMs (Table 1), indicating structural heterogeneity. Nonetheless, interesting SAR trends were detected. From 2006 to 2011, added inhibitors often represented new analogue series (MMSs), resulting in a gradual increase in the number of new SARMs (magenta) during this period. In 2011 and especially 2013, a larger number of expanded SARMs (blue) was observed, indicating follow-up investigations on existing series. Between 2011 and 2013, a shift of expanded and new SARMs toward the upper right quadrant of the plots was observed, revealing overall promising SAR progression.

The set of C−C chemokine receptor type 2 ligands in Figure 3B was much larger (836 compounds) than the cytochrome P450 11B1 inhibitor set and ultimately yielded 1365 SARMs (resulting in high-density scatterplots). Between 2006 and 2008, a shift of the SARM distributions toward the right of the plots was observed. During subsequent years, the distributions became increasingly dominated by a large number of new SARMs with high median potency (in addition, SARM expansion was also observed). Thus, many novel series containing highly potent compounds became available, reflecting successful compound design efforts. A different picture emerged for ligands of melanin-concentrating hormone receptor 1 in Figure 3C, the largest data set (990 compounds) producing 1086 SARMs. In 2005, the distribution was dominated by new SARMs (resulting from structurally novel compounds not available during the preceding year). In 2006, many SARMs were expanded, reflecting follow-up chemistry efforts, and the distribution shifted toward higher potency and discontinuity scores, indicating SAR progression. However, during 2007 and 2008, the number of new and expanded

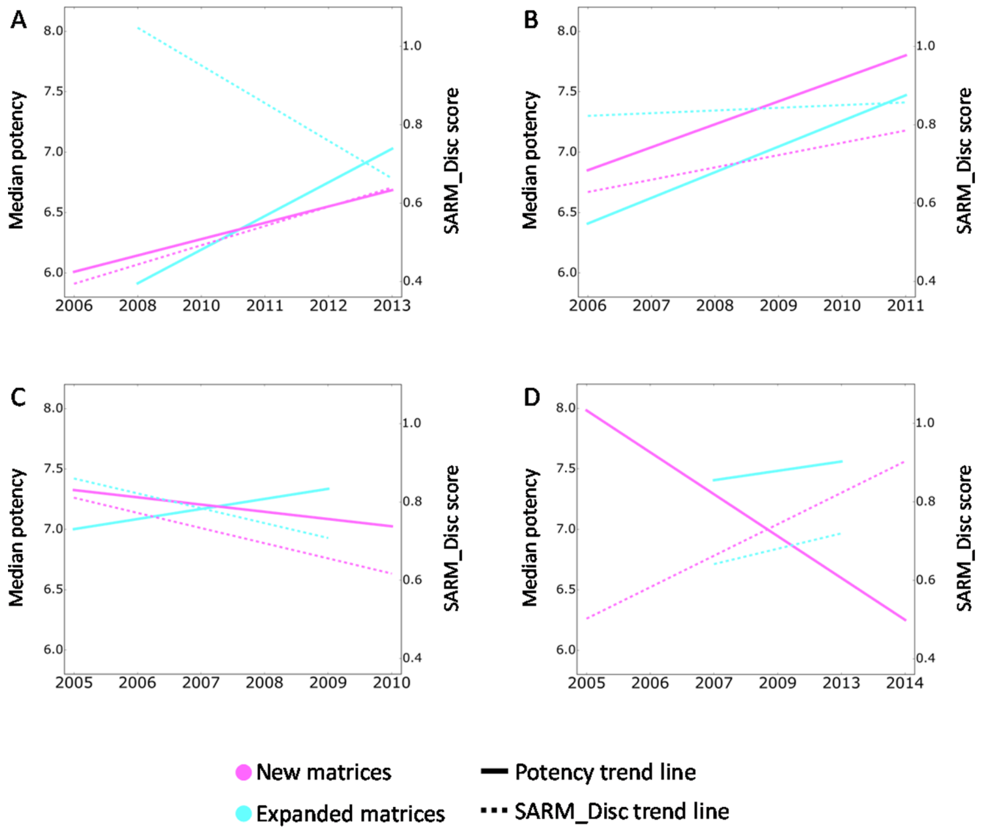


Figure 4. Trend plots of expanded and new indicator SARMs according to Figure 2B derived from the data distributions in Figure 3. (A) Cytochrome P450 11B1 inhibitors, (B) C−C chemokine receptor type 2 ligands, (C) melanin-concentrating hormone receptor 1 ligands, and (D) caspase-7 inhibitors. Trend lines separately monitor the development of median potency and SARM_Disc scores over time for a given category of indicator SARMs.

SARMs declined, indicating reduced chemistry efforts. Another boost in novel active compounds was detected in 2009, which further increased median potency. However, there was essentially no matrix expansion in 2010, and the number of new SARMs also declined again. Hence, in this case, different intervals of strong and weak SAR progression were detected. Figure 3D monitors the smallest of the four data sets, consisting of 232 inhibitors of caspase-7, that yielded a total of only 125 SARMs. Although the number of SARMs was small in this case, their temporal distributions revealed an obvious trend. During 2005 and 2006, a limited number of inhibitors and SARMs became available, and expanded SARMs were first detected in 2007. However, between 2009 and 2014, an increasing number of SARMs was found to map to the upper left quadrant of the plots, characterized by the presence of low median potency and high discontinuity, resulting from the addition of more and more weakly potent compounds to a small number of highly potent ones. Thus, in this case, negative SAR progression was observed.

Figure 4 reports trend plots for new and expanded SARMs generated from the distributions in Figure 3. Especially for very large SARM ensembles, trend lines that separately monitor potency and discontinuity help to better understand characteristics of SAR progression, although they are only approximate. Figure 4A confirms the conclusions drawn from SARM distribution analysis for the cytochrome P450 11B1 inhibitor set. The median potency and discontinuity score of new SARMs were increasing, and potency of expanded SARMs also increased. The only exception to overall positive SAR progression was the observed decrease in discontinuity of expanded SARMs, which likely resulted from the increasing number of analogues of existing series having comparable potency. Furthermore, Figure 4B also reveals a clear example of positive SAR progression, consistent with SARM distribution analysis, for the large set of C−C chemokine receptor type 2 ligands. In this case, median potency and discontinuity increased for all SARMs or remained essentially constant at a high level (i.e., discontinuity of expanded SARMs). Figure 4C reflects overall limited SAR progression for the set of melanin-concentrating hormone receptor 1 ligands, as discussed, and Figure 4D displays negative trends for caspase-7 inhibitors. Here, a strong decline of median potency was detected for new SARMs, which was accompanied by an increase in discontinuity. Although this observation might be puzzling at a first glance, it can be easily rationalized as resulting from the presence of analogues with decreasing potency in SARMs also containing highly potent compounds. Furthermore, for a small number of expanded SARMs, potency increased only slightly and discontinuity remained at a low level.

Taken together, temporal distribution analysis of indicator SARMs from exemplary target-based compound sets evolving over time detected clear differences in SAR progression, hence providing support for the underlying methodological concept. Next, actual LO data sets originating from drug discovery were investigated. Such data sets are currently not available in the public domain.

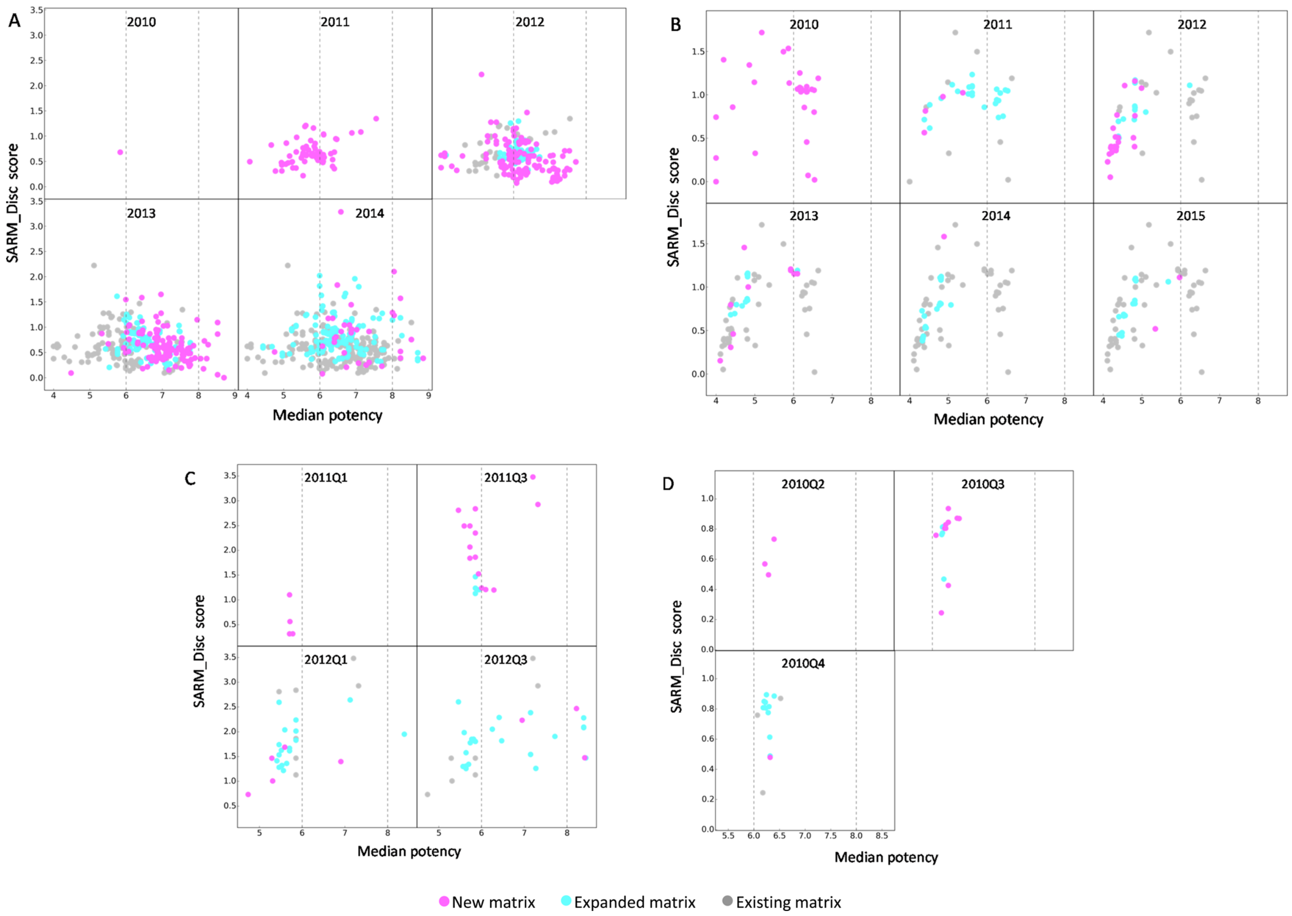


Figure 5. Indicator SARM distributions over a time course for LO sets. Scatterplots are shown for two LO data sets that were assembled from Pfizer project team data on the basis of project progression information. (A) Neurodegenerative target, series 1, (B) neurodegenerative target, series 2, (C) inflammation target, series 1, and (D) inflammation target, series 2. Compound and SARM statistics for the monitored time periods are provided in Table 2. Series 1 in (A) and (C) represented successful project progressions from which compounds were nominated as candidates for preclinical studies. By contrast, series 2 in (B) and (D) represented unsuccessful project progressions from which no compounds were nominated.

LO Data Sets. Two LO sets from different Pfizer drug discovery projects were investigated. Each project team pursued two different chemical series per target. In each case, one of the series was considered to be successful because the project team was able to nominate preclinical candidate(s) from this series, and the other series was unsuccessful, yielding no candidate compounds. Table 2 provides a description of these LO sets.

Figure 5 shows the distribution of indicator SARMs obtained over a period of 4 to 5 years for the neurodegenerative target and 3 to 7 quarters for the inflammation target. Figure 5A monitors the SAR progression of series 1 of the neurodegenerative project. This set ultimately yielded 672 SARMs for 431 compounds (Table 2), indicating structural homogeneity. In 2010, LO efforts on this series started with 10 analogues active in the micromolar range contained in a single SARM. Figure 5A reveals that there was consistent positive SAR progression for series 1. Starting in 2012, new and expanded SARMs were detected, and there were clear breakthroughs in 2013 and 2014, yielding highly potent compounds in increasingly informative SAR environments. On the basis of SAR monitoring, LO on series 1 was a highly promising project, consistent with its ultimate success. Similar trends were not observed for series 2 in Figure 5B, although there was much more compound and SAR information available initially than that for series 1. LO efforts on series 2 started with 46 compounds, and a total of 125 inhibitors were evaluated over a period of 6 years. However, the project team was unable to break a potency barrier with this chemical series. Although matrix expansion occurred during the first 3 years, no notable SAR progression was detected, and in 2014, it was evident that the LO project faced a roadblock.

The comparably small series of inflammation inhibitors in Figure 5C,D with, ultimately, 88 and 78 compounds, respectively, also exhibited rather different SAR progression. Series 1 in Figure 5C displayed very positive SAR trends with significantly increasing SAR information content and compound potency already detectable during the first two time intervals. By contrast, very little SAR progression was observed for series 2 in Figure 5D from the second to the third quarter of 2010, but no further progression was observed during the fourth quarter. Thus, SAR monitoring contrasts these two series of inflammation inhibitors, and it is easy to reconcile why series 1 was ultimately successful and series 2 was not.

The trend plots for these LO sets in Figure 6 strongly support conclusions drawn from indicator SARM distribution analysis. The successful series 1 of neurodegeneration inhibitors in Figure 6A and inflammation inhibitors in Figure 6C displayed an increase in all trend lines for new and expanded

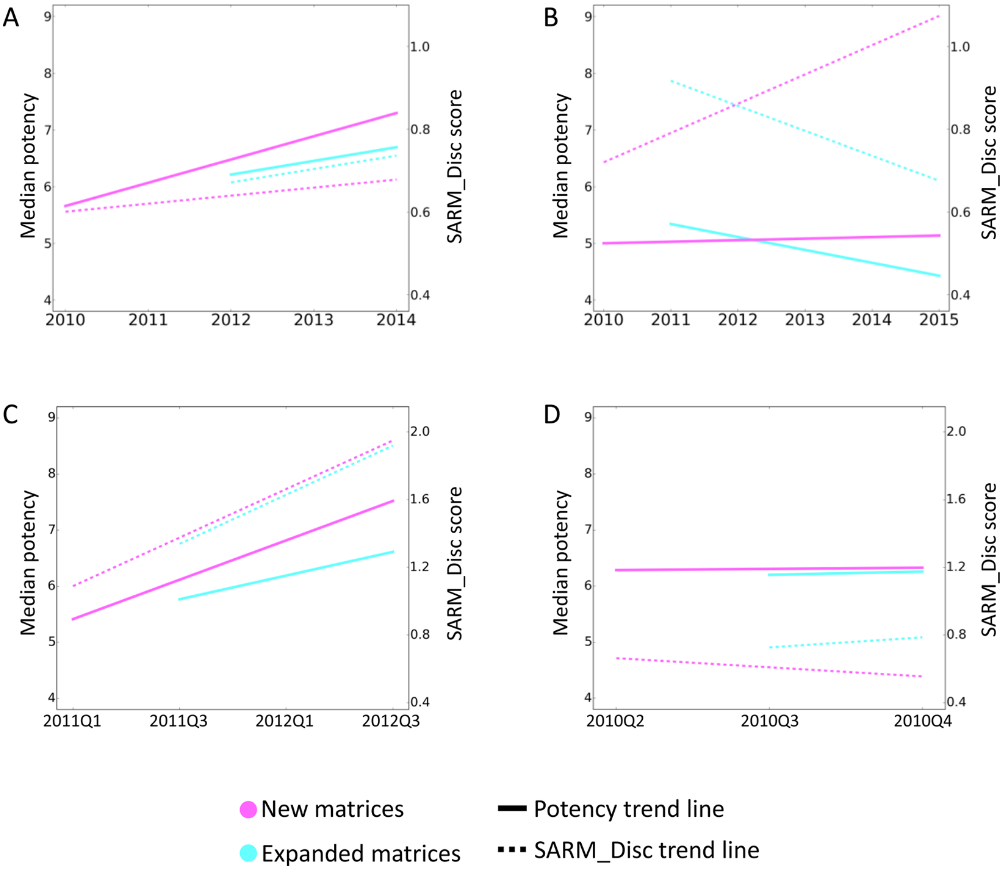


Figure 6. Trend plots for LO sets showing expanded and new indicator SARMs derived from the data distributions in Figure 5. (A) Neurodegenerative target, series 1, (B) neurodegenerative target, series 2, (C) inflammation target, series 1, and (D) inflammation target, series 2. Trend lines separately monitor the development of median potency and SARM_Disc scores over time for a given category of indicator SARMs. Series 1 in (A) and (C) represented successful chemical series and displayed positive SAR progression with an increase in both median potency and SARM_Disc scores. Series 2 in (B) and (D) represented unsuccessful chemical series, which displayed negative SAR progression for expanded SARMs with a decrease in median potency and SARM_Disc scores and essentially flat SARs for new SARMs.

matrices. By contrast, the unsuccessful series 2 of neurodegeneration inhibitors in Figure 6B was characterized by decreasing trend lines for expanded matrices, reflecting negative SAR progression of close-in analoging attempts and diverging trend lines for new matrices, with an increase in SAR information content resulting from the addition of new but only weakly potent compounds that could not be further optimized. Moreover, the series 2 of inflammation inhibitors in Figure 6D displayed essentially flat SAR characteristics throughout.

On the basis of the comparisons reported in Figures 5 and 6, successful LO series of neurodegeneration and inflammation inhibitors were clearly distinguished from unsuccessful series. Analysis of indicator SARM distributions would have made it possible to predict the lack of SAR progression for the latter series during the course of LO.

■ CONCLUSIONS

Lead optimization is a largely hypothesis-driven process that depends mainly on medicinal chemistry experience and intuition. Only few efforts have thus far been made to rationalize this process and assess LO progress. Efforts in this direction are highly desirable to support decision making because it is very difficult to predict the ultimate outcome of LO campaigns and control the number of compounds to be evaluated before meaningful conclusions can be reached. In this study, we have introduced a computational framework to monitor the progression of SAR information content during LO over a time course. The SAR matrix data structure, which was originally developed for a completely different purpose, i.e., the elucidation of SAR patterns in related analogue series and compound prediction, was adapted as a diagnostic tool to evaluate SAR progression. This was accomplished by generation of SARM ensembles for compound sets evolving over time, classification of SARMs based on the compounds they contain, and characterization of their SAR information content. SAR information contained in individual SARMs was quantified on the basis of a newly introduced matrix discontinuity score combined with median potency calculations. Characteristic shifts of SARM ensembles in scatter plots were found to indicate positive, neutral, or negative SAR progression and revealed significant differences between target-based compound sets. Analysis of SARM distributions was complemented by trend plots designed to summarize SAR progression over time. Our proof-of-concept investigations show that SARM ensembles are capable of detecting differences in SAR progression in compound sets of distinct composition. As a diagnostic tool, they can be used to distinguish SAR progression from redundancy, i.e., when increasing numbers of

DOI: 10.1021/acs.jmedchem.5b01428

compounds are made that do not add novel SAR information or further improve potency. Application of the approach to actual LO sets from drug discovery projects revealed very clear SAR trends over time for series that were ultimately successful or unsuccessful. Such insights are valuable in project decision making. Taken together, the results reported herein suggest that indicator SARMs should merit further investigation in LO assessment. Since the SARM data structure can be easily annotated with different molecular properties, multiple parameters can be monitored.

■ AUTHOR INFORMATION

Corresponding Author

*Tel: +49-228-2699-306. Fax: +49-228-2699-341. E-mail: bajorath@bit.uni-bonn.de.

Author Contributions

‖S.K. and A.d.l.V.d.L. contributed equally to this work.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

The coauthors of the University of Bonn would like to thank OpenEye for providing an academic license. For the two Pfizer LO sets, a waiver on ACS data deposition requirements has been granted.

■ ABBREVIATIONS USED

LO, lead optimization; MMP, matched molecular pair; MMS, matching molecular series; SAR, structure−activity relationship; SARM, SAR matrix

■ REFERENCES

(1) *The Practice of Medicinal Chemistry*, 3rd ed.; Wermuth, C. G., Ed.; Academic Press: Boston, MA, 2008.

(2) Nicolaou, C. A.; Brown, N.; Pattichis, C. S. Molecular Optimization Using Computational Multi-Objective Methods. *Curr. Opin. Drug. Discovery Develop.* **2007**, *10*, 316−324.

(3) Segall, M. Advances in Multi-Parameter Optimization Methods for *De Novo* Drug Design. *Expert Opin. Drug Discovery* **2014**, *9*, 803−817.

(4) Iyer, P.; Hu, Y.; Bajorath, J. SAR Monitoring of Evolving Compound Data Sets Using Activity Landscapes. *J. Chem. Inf. Model.* **2011**, *51*, 532−540.

(5) Maynard, A. T.; Roberts, C. D. Quantifying, Visualizing, and Monitoring Lead Optimization. *J. Med. Chem.* **2015**, DOI: 10.1021/acs.jmedchem.5b00948.

(6) Munson, M.; Lieberman, H.; Tserlin, E.; Rocnik, J.; Ge, J.; Fitzgerald, M.; Patel, V.; Garcia-Echeverria, C. Lead Optimization Attrition Analysis (LOAA): A Novel and General Methodology for Medicinal Chemistry. *Drug Discovery Today* **2015**, *20*, 978−987.

(7) Wassermann, A. M.; Haebel, P.; Weskamp, N.; Bajorath, J. SAR Matrices: Automated Extraction of Information-Rich SAR Tables from Large Compound Data Sets. *J. Chem. Inf. Model.* **2012**, *52*, 1769−1776.

(8) Gupta-Ostermann, D.; Bajorath, J. The 'SAR Matrix' Method and its Extensions for Applications in Medicinal Chemistry and Chemogenomics. *F1000Research* **2014**, *3*, 113.

(9) Kenny, P. W.; Sadowski, J. Structure Modification in Chemical Databases. In *Chemoinformatics in Drug Discovery*; Oprea, T. I., Ed.; Wiley-VCH: Weinheim, Germany, 2005; pp 271−285.

(10) Hussain, J.; Rea, C. Computationally Efficient Algorithm to Identify Matched Molecular Pairs (MMPs) in Large Data Sets. *J. Chem. Inf. Model.* **2010**, *50*, 339−348.

(11) Wawer, M.; Bajorath, J. Local Structural Changes, Global Data Views: Graphical Substructure−Activity Relationship Trailing. *J. Med. Chem.* **2011**, *54*, 2944−2951.

(12) *OEChem TK*; OpenEye Scientific Software, Inc.: Santa Fe, NM, 2012.

(13) Peltason, L.; Bajorath, J. SAR Index: Quantifying the Nature of Structure−Activity Relationships. *J. Med. Chem.* **2007**, *50*, 5571−5578.

(14) Gaulton, A.; Bellis, L. J.; Bento, A. P.; Chambers, J.; Davies, M.; Hersey, A.; Light, Y.; McGlinchey, S.; Michalovich, D.; Al-Lazikani, B.; Overington, J. P. ChEMBL: A Large-Scale Bioactivity Database for Drug Discovery. *Nucleic Acids Res.* **2012**, *40*, D1100−D1107.

(15) Shanmugasundaram, V.; Zhang, L.; Kayastha, S.; de la Vega de León, A.; Dimova, D.; Bajorath, J. Data Sets for SAR Progression Analysis. *Zenodo* **2015**, DOI: 10.5281/zenodo.32794.

Summary

SARMs, originally introduced to analyze structural relationships between analogs and identify SAR patterns in data sets, have been applied in our study for the purpose of determining the progression of SAR information in lead optimization projects. Modifications were introduced to SARMs to account for their SAR information content. Four evolving data sets were assembled from ChEMBL and two drug discovery projects at Pfizer. Each of the lead optimization sets from Pfizer comprised two series, a successful one for which a clinical candidate compound was proposed and an unsuccessful one with none of the compounds qualified as clinical candidate. SARM ensembles were obtained for each of these sets at individual time points and then classified into three categories: existing, expanded and new SARMs. An existing SARM was not modified by addition of new compounds and was present in the previous time point. An expanded SARM resulted when new structural analogs were added to an already existing series. New SARMs arose when completely new analog series were formed. SAR information in a given indicator SARM was quantified by calculating meadian potency of all compounds contained in it and SARM-based SAR discontinuity score. Progress of SAR information during lead optimization corresponds to generation of SARMs with high SAR discontinuity score and median potency during the time course of the project. These two scores were used to graphically monitor SAR progression in scatterplots generated for each time point. Characteristic difference in SAR progression was observed in two exemplary data sets from ChEMBL. Further, clear distinctions could be made between successful and unsuccessful series obtained from lead optimization projects. Taken together, we demonstrate that indicator SARMs can be utilized as a diagnostic tool to monitor SAR progression over time. The approach can be used to distinguish SAR progression from redundancy. These insights are valuable and should support the process of decision making in lead optimization projects.

My major contributions to this work include pre-processing SAR matrix files for temporal analysis and generating the scatterplot distribution of classified SARMs for individual time points.

Chapter 9

Conclusions and Perspectives

Computational exploration of chemical spaces is multifaceted and several approaches already exist that have been applied to rationalize the chemical universe and extract meaningful patterns. The current chemical universe comprises a vast amount of compounds covering a wide range of structural features and molecular properties. It therefore represents a valuable resource to derive further insight into composition and distribution of compounds in the space as well as relationships between their features. To tackle the challenges faced during analysis of the huge amount of data in chemical spaces, a variety of computational methods have been developed. Among these are methods for SAR analysis that are of prime interest in medicinal chemistry and chemoinformatics. This thesis work is dedicated to developing new computational approaches for large-scale chemical space analysis, specifically focusing on visualization and predictions of SAR features. Additionally, methods to systematically extract and analyze structural patterns from target-specific data sets from publicly available compound repositories have been presented.

The first representative study (*Chapter 2*) provided an extensive scaffold survey from bioactive compounds reported so far in literature. The major goal of this analysis was to systematically capture structurally distinct scaffold and CSK combinations to assess the structural diversity among bioactive compounds in ChEMBL. Moreover, the potency range distribution of compounds representing such structurally diverse combinations of scaffolds and CSKs were determined. Our analysis has revealed an abundance of structurally diverse and highly potent pairs of compounds targeting many pharmaceutically relevant proteins. This finding supports future research directed towards identifying novel structural compound classes as an alternative to currently available compound series. GTM-based extraction of privileged structural motifs for major target families in ChEMBL was the aim of our next study (*Chapter 3*). This approach extended traditional scaffold-based definition of privileged substructures and demonstrated the application of GTM to prioritize compound clusters and subsequently detect structural motifs privileged towards specific target superfamilies from those clusters. GTM-based clustering of compounds provided the evidence of its ability to delineate regions of chemical space that were preferentially occupied by compounds belonging to individual target superfamilies. The substructures obtained from our methodology were subjected to careful visual inspection and identified as motifs present in specifically active compounds as well as approved drugs.

Large-scale SAR analysis are often accomplished by employing visualization techniques. A two-component visualization tool combining coordinate-based GTM and coordinate-free CSN representations was introduced (*Chapter 4*). This combined approach facilitated smooth transition from global views of SAR landscape to local views of specific molecular communities characterized by interesting SAR patterns. The prominent feature of our global-local visualization technique is its ability to handle large and heterogeneous data sets. Prediction of activity cliff is one of the least explored territories of SAR analysis. We have applied machine learning methods to derive computational models for activity cliff prediction (*Chapter 5*). Alternative representations of MMPs, based on condensed graph of reactions and descriptor recombination were used as an input to support vector machine models. Successful classification and regression models were obtained for predicting whether a given MMP formed an activity cliff and signed potency difference between MMP partners, respectively.

Ionization state of compounds play a crucial role in their biological activity as well as other important drug relevant properties. A large-scale analysis of calculated ionization states of publicly available bioactive compounds was performed emphasizing primarily on their charge state distribution across different potency ranges and chemical neighborhoods (*Chapter 6*). Global distribution of charge states showed that majority of bioactive compounds were either neutral or basic under physiological condition and within individual target sets, families and superfamilies notable differences in preferred ionization states were observed. Majority of chemical neighborhoods of bioactive compounds retained their charge sates and only in one-third of the neighborhoods they varied. Highly potent compounds in most of the target sets showed preference towards a particular ionization state.

In the remaining part of the thesis, methodologies developed to address major challenges in compound optimization have been highlighted. Computational multi-property optimization algorithms that are applied to reach a balance between various drug-relevant properties of compounds often produce more than one equally optimal solution. A visualization framework employing star and parallel coordinate representations was developed to complement multi-objective optimization tasks (*Chapter 7*). Star coordinates allowed visualization of optimized projections, whereby compound distribution in multi-dimensional property space is projected onto 2D. We demonstrated that start coordinates aid in differentiating numerically equivalent solutions arising from similar or different property settings. Lastly, we utilized SAR matrix data structure for determining the SAR progression in lead optimization data sets (*Chapter 8*). SAR matrices were generated for evolving data sets, each matrix was then scored based on SAR discontinuity and median potency of their compounds. Graphical analysis of plots generated using these two scores were indications for positive, negative or neutral SAR progression.

Taken together, this thesis work has introduced new methodological developments for systematic analysis of chemical spaces, with major focus on large-scale SAR analysis. Additionally, new methods of chemical space visualization have been introduced that allow prioritization of desired subspaces to aid compound design and optimization efforts in drug discovery.

Even though a substantial amount of insights have been gained through methods developed so far to analyze the vast chemical space, we believe that further extensions and enhancement of the existing methodologies and introduction of novel techniques are nevertheless necessary. In the following, we will discuss about the major limitations of existing methods as well as an outlook for future perspectives.

While thousands of molecular descriptors are available for visualizing compounds in high-dimensional chemical space, analyzing, and modeling their SAR characteristics, they offer only limited interpretation from medicinal chemistry point of view. Moreover, most of the graphical SAR analysis methods are descriptive in nature. Using such approaches, it is rather difficult to address the most common situation faced in medicinal chemistry practice, i.e., decision about the next compound to be synthesized. Despite chemically intuitive nature of substructure-based graphical SAR analysis approaches, only very few of them could directly assist in compound design. Since, the main goal of chemoinformatics is implementing computational methods to solve chemical problems, specifically in our case, to tackle the challenges of drug discovery, it is important that the methods could be easily applied in practice by medicinal chemists. One of the possibilities to do so would be combined approaches that integrate the power of molecular descriptors in handling large amount of SAR data with chemical intuitiveness of substructure-based methods. The two-layered SAR visualization tool based on GTM and CSN methodologies presented in this thesis work is an example of such approach. The feature of CSN that allows to interactively navigate through chemical space by focusing on individual compounds represented as nodes that can be associated with their structures is relatively more interesting from a chemist's perspective compared to mathematical models with black-box characteristics. However, such SAR networks are still considered to be complicated for medicinal chemists who are accustomed to work with R-group table like views. On the other hand, given the increasing amount of SAR data, it would be indispensable for chemists to use graphical SAR analysis methods. Therefore, future direction of SAR research should implement tools that are practical, interpretable and easily accessible by medicinal chemists.

References

[1] Kubinyi, H. Privileged structures and analogue-based drug discovery. In *Analogue-based Drug Discovery*; Wiley-VCH Verlag GmbH & Co. KGaA: Weinheim, 2006; pp 53–68.

[2] Martin, Y. C.; Muchmore, S. Beyond QSAR: lead hopping to different structures. *Molecular Informatics* **2009**, *28*, 797–801.

[3] Bemis, G. W.; Murcko, M. A. The properties of known drugs. 1. Molecular frameworks. *Journal of medicinal chemistry* **1996**, *39*, 2887–2893.

[4] Gaulton, A.; Bellis, L. J.; Bento, A. P.; Chambers, J.; Davies, M.; Hersey, A.; Light, Y.; McGlinchey, S.; Michalovich, D.; Al-Lazikani, B.; Overington, J. P. ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic acids research* **2012**, *40*, D1100–D1107.

[5] Bishop, C. M.; Svensén, M.; Williams, C. K. GTM: The generative topographic mapping. *Neural computation* **1998**, *10*, 215–234.

[6] Sidorov, P.; Gaspar, H.; Marcou, G.; Varnek, A.; Horvath, D. Mappability of drug-like space: towards a polypharmacologically competent map of drug-relevant compounds. *Journal of computer-aided molecular design* **2015**, *29*, 1087–1108.

[7] Stumpfe, D.; Bajorath, J. Exploring activity cliffs in medicinal chemistry: miniperspective. *Journal of medicinal chemistry* **2012**, *55*, 2932–2942.

[8] Brown, F. K. Chemoinformatics: what is it and how does it impact drug discovery. *Annual reports in medicinal chemistry* **1998**, *33*, 375–384.

[9] Engel, T. Basic overview of chemoinformatics. *Journal of chemical information and modeling* **2006**, *46*, 2267–2277.

[10] Bunin, B. A.; Siesel, B.; Morales, G.; Bajorath, J. Chemoinformatics Theory. In *Chemoinformatics: Theory, Practice, & Products*; Springer, 2007; pp 1–49.

[11] Hann, M.; Green, R. Chemoinformatics-a new name for an old problem? *Current Opinion in Chemical Biology* **1999**, *3*, 379–383.

[12] Bajorath, J. Understanding chemoinformatics: a unifying approach. *Drug discovery today* **2004**, *9*, 13–14.

[13] Hughes, J. P.; Rees, S.; Kalindjian, S. B.; Philpott, K. L. Principles of early drug discovery. *British journal of pharmacology* **2011**, *162*, 1239–1249.

[14] Bleicher, K. H.; Böhm, H.-J.; Müller, K.; Alanine, A. I. Hit and lead generation: beyond high-throughput screening. *Nature reviews Drug discovery* **2003**, *2*, 369–378.

[15] Greaves, P.; Williams, A.; Eve, M. First dose of potential new medicines to humans: how animals help. *Nature Reviews Drug Discovery* **2004**, *3*, 226–236.

[16] Oprea, T. I. Chemoinformatics and the quest for leads in drug discovery. In *Handbook of Chemoinformatics: From Data to Knowledge in 4 Volumes*; Wiley Online Library, 2003; Vol. 4; pp 1508–1531.

[17] Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *Journal of chemical information and computer sciences* **1988**, *28*, 31–36.

[18] Heller, S. R.; McNaught, A.; Pletnev, I.; Stein, S.; Tchekhovskoi, D. InChI, the IUPAC international chemical identifier. *Journal of cheminformatics* **2015**, *7*, 23.

[19] Schuffenhauer, A.; Varin, T. Rule-Based Classification of Chemical Structures by Scaffold. *Molecular informatics* **2011**, *30*, 646–664.

[20] Hu, Y.; Stumpfe, D.; Bajorath, J. Lessons learned from molecular scaffold analysis. *Journal of chemical information and modeling* **2011**, *51*, 1742–1753.

[21] Xu, Y.-j.; Johnson, M. Algorithm for naming molecular equivalence classes represented by labeled pseudographs. *Journal of chemical information and computer sciences* **2001**, *41*, 181–185.

[22] Peter, W. K.; Jens, S. Structure modification in chemical databases. *Chemoinformatics in drug discovery* **2006**, *23*.

[23] Dimova, D.; Stumpfe, D.; Bajorath, J. Quantifying the fingerprint descriptor dependence of structure–activity relationship information on a large scale. *Journal of chemical information and modeling* **2013**, *53*, 2275–2281.

[24] Hu, X.; Hu, Y.; Vogt, M.; Stumpfe, D.; Bajorath, J. MMP-cliffs: systematic identification of activity cliffs on the basis of matched molecular pairs. *Journal of chemical information and modeling* **2012**, *52*, 1138–1145.

[25] Leach, A. G.; Jones, H. D.; Cosgrove, D. A.; Kenny, P. W.; Ruston, L.; MacFaul, P.; Wood, J. M.; Colclough, N.; Law, B. Matched molecular pairs as a guide in the optimization of pharmaceutical properties; a study of aqueous solubility, plasma protein binding and oral exposure. *Journal of medicinal chemistry* **2006**, *49*, 6672–6682.

[26] Gleeson, P.; Bravi, G.; Modi, S.; Lowe, D. ADMET rules of thumb II: A comparison of the effects of common substituents on a range of ADMET parameters. *Bioorganic & medicinal chemistry* **2009**, *17*, 5906–5919.

[27] Haubertin, D. Y.; Bruneau, P. A database of historically-observed chemical replacements. *Journal of chemical information and modeling* **2007**, *47*, 1294–1302.

[28] Sheridan, R. P. The most common chemical replacements in drug-like compounds. *Journal of chemical information and computer sciences* **2002**, *42*, 103–108.

[29] Raymond, J. W.; Watson, I. A.; Mahoui, A. Rationalizing lead optimization by associating quantitative relevance with molecular structure modification. *Journal of chemical information and modeling* **2009**, *49*, 1952–1962.

[30] Hussain, J.; Rea, C. Computationally efficient algorithm to identify matched molecular pairs (MMPs) in large data sets. *Journal of chemical information and modeling* **2010**, *50*, 339–348.

[31] Brugger, W.; Stuper, A.; Jurs, P. C. Generation of descriptors from molecular structures. *Journal of Chemical Information and Computer Sciences* **1976**, *16*, 105–110.

[32] Glen, R.; Rose, V. Computer program suite for the calculation, storage and manipulation of molecular property and activity descriptors. *Journal of Molecular Graphics* **1987**, *5*, 79–86.

[33] Todeschini, R.; Consonni, V. *Handbook of molecular descriptors*; John Wiley & Sons, 2008; Vol. 11.

[34] Laboratoire de Chemoinformatique, U. o. S. ISIDA Fragmentor Software. `http://infochim.u-strasbg.fr/spip.php?rubrique41`.

[35] Lacomme, P.; Prins, C.; Sevaux, M. *Algorithmes de graphes*; Eyrolles Paris, 2003; Vol. 28.

[36] Varnek, A.; Fourches, D.; Hoonakker, F.; Solov'ev, V. P. Substructural fragments: an universal language to encode reactions, molecular and supramolecular structures. *Journal of computer-aided molecular design* **2005**, *19*, 693–703.

[37] Varnek, A.; Fourches, D.; Horvath, D.; Klimchuk, O.; Gaudin, C.; Vayer, P.; Solov'ev, V.; Hoonakker, F.; Tetko, I. V.; Marcou, G. ISIDA-Platform for virtual screening based on fragment and pharmacophoric descriptors. *Current Computer-Aided Drug Design* **2008**, *4*, 191.

[38] Bonachéra, F.; Parent, B.; Barbosa, F.; Froloff, N.; Horvath, D. Fuzzy tricentric pharmacophore fingerprints. 1. Topological fuzzy pharmacophore triplets and adapted molecular similarity scoring schemes. *Journal of chemical information and modeling* **2006**, *46*, 2457–2477.

[39] Bonachéra, F.; Horvath, D. Fuzzy tricentric pharmacophore fingerprints. 2. Application of topological fuzzy pharmacophore triplets in quantitative structure- activity relationships. *Journal of chemical information and modeling* **2008**, *48*, 409–425.

[40] Ruggiu, F.; Marcou, G.; Varnek, A.; Horvath, D. ISIDA Property-Labelled Fragment Descriptors. *Molecular Informatics* **2010**, *29*, 855–868.

[41] Chemical Computing Group Inc. Montreal, H. R. C. Molecular Operating Environment(MOE).

[42] Lin, A. QuaSAR-descriptors. *Chemical Computing Group Inc. Montreal, H3A 2R7 Canada* **2002**,

[43] Symyx Software, C. U., San Ramon MACCS Structural Keys. 2002.

[44] James, C.; Weininger, D.; Delany, J. Daylight theory manual daylight version 4.82. `http://www.daylight.com/dayhtml/doc/theory/`, 2003.

[45] Leach, A. R.; Gillet, V. J. *An introduction to chemoinformatics*; Springer Science & Business Media, 2007.

[46] Bender, A.; Glen, R. C. Molecular similarity: a key technique in molecular informatics. *Organic & biomolecular chemistry* **2004**, *2*, 3204–3218.

[47] Bender, A.; Mussa, H. Y.; Glen, R. C.; Reiling, S. Similarity searching of chemical databases using atom environment descriptors (MOLPRINT 2D): evaluation of performance. *Journal of chemical information and computer sciences* **2004**, *44*, 1708–1718.

[48] Willett, P. Similarity-based virtual screening using 2D fingerprints. *Drug discovery today* **2006**, *11*, 1046–1053.

[49] Stumpfe, D.; Bajorath, J. Similarity searching. *Wiley Interdisciplinary Reviews: Computational Molecular Science* **2011**, *1*, 260–282.

[50] Takaoka, Y.; Endo, Y.; Yamanobe, S.; Kakinuma, H.; Okubo, T.; Shimazaki, Y.; Ota, T.; Sumiya, S.; Yoshikawa, K. Development of a method for evaluating drug-likeness and ease of synthesis using a data set in which compounds are assigned scores based on chemists' intuition. *Journal of chemical information and computer sciences* **2003**, *43*, 1269–1275.

[51] Lajiness, M. S.; Maggiora, G. M.; Shanmugasundaram, V. Assessment of the consistency of medicinal chemists in reviewing sets of compounds. *Journal of medicinal chemistry* **2004**, *47*, 4891–4896.

[52] Kutchukian, P. S.; Vasilyeva, N. Y.; Xu, J.; Lindvall, M. K.; Dillon, M. P.; Glick, M.; Coley, J. D.; Brooijmans, N. Inside the mind of a medicinal chemist: the role of human bias in compound prioritization during drug discovery. *PloS one* **2012**, *7*, e48476.

[53] Maggiora, G.; Vogt, M.; Stumpfe, D.; Bajorath, J. Molecular similarity in medicinal chemistry: miniperspective. *Journal of medicinal chemistry* **2013**, *57*, 3186–3204.

[54] Willett, P.; Barnard, J. M.; Downs, G. M. Chemical similarity searching. *Journal of chemical information and computer sciences* **1998**, *38*, 983–996.

[55] Kubinyi, H. QSAR in drug design. In *Handbook of Chemoinformatics: From Data to Knowledge in 4 Volumes*; Wiley Online Library, 2003; Vol. 4; pp 1532–1554.

[56] Esposito, E. X.; Hopfinger, A. J.; Madura, J. D. Methods for applying the quantitative structure-activity relationship paradigm. *Chemoinformatics: Concepts, Methods, and Tools for Drug Discovery* **2004**, 131–213.

[57] Johnson, M. A.; Maggiora, G. M. *Concepts and applications of molecular similarity*; Wiley, 1990.

[58] Wassermann, A. M.; Wawer, M.; Bajorath, J. Activity landscape representations for structure- activity relationship analysis. *Journal of medicinal chemistry* **2010**, *53*, 8209–8223.

[59] Bajorath, J.; Peltason, L.; Wawer, M.; Guha, R.; Lajiness, M. S.; Van Drie, J. H. Navigating structure–activity landscapes. *Drug Discovery Today* **2009**, *14*, 698–705.

[60] Bajorath, J. Large-scale SAR analysis. *Drug Discovery Today: Technologies* **2013**, *10*, e419–e426.

[61] Shanmugasundaram, V.; Maggiora, G. Characterizing activity landscapes using an information theoretic approach. 222nd ACS National Meeting. 2001; pp 26–30.

[62] Tetko, I. V.; Kovalishyn, V. V.; Livingstone, D. J. Volume learning algorithm artificial neural networks for 3D QSAR studies. *Journal of medicinal chemistry* **2001**, *44*, 2411–2420.

[63] Livingstone, D. J.; Manallack, D. T. Neural networks in 3D QSAR. *Molecular Informatics* **2003**, *22*, 510–518.

[64] Guha, R.; Jurs, P. C. Interpreting computational neural network QSAR models: a measure of descriptor importance. *Journal of chemical information and modeling* **2005**, *45*, 800–806.

[65] Sussman, N.; Arena, V.; Yu, S.; Mazumdar, S.; Thampatty, B. Decision tree SAR models for developmental toxicity based on an FDA/TERIS database. *SAR and QSAR in Environmental Research* **2003**, *14*, 83–96.

[66] Zhang, Q.-Y.; Aires-de Sousa, J. Random forest prediction of mutagenicity from empirical physicochemical descriptors. *Journal of chemical information and modeling* **2007**, *47*, 1–8.

[67] Cortes, C.; Vapnik, V. Support-vector networks. *Machine learning* **1995**, *20*, 273–297.

[68] Burbidge, R.; Trotter, M.; Buxton, B.; Holden, S. Drug design by machine learning: support vector machines for pharmaceutical data analysis. *Computers & chemistry* **2001**, *26*, 5–14.

[69] Byvatov, E.; Fechner, U.; Sadowski, J.; Schneider, G. Comparison of support vector machine and artificial neural network systems for drug/nondrug classification. *Journal of chemical information and computer sciences* **2003**, *43*, 1882–1889.

[70] Kotsiantis, S. B.; Zaharakis, I.; Pintelas, P. Supervised machine learning: A review of classification techniques. 2007.

[71] Vapnik, V. *The nature of statistical learning theory*; Springer science & business media, 2013.

[72] Burges, C. J. A tutorial on support vector machines for pattern recognition. *Data mining and knowledge discovery* **1998**, *2*, 121–167.

[73] Smola, A. J.; Schölkopf, B. A tutorial on support vector regression. *Statistics and computing* **2004**, *14*, 199–222.

[74] Dobson, C. M. Chemical space and biology. *Nature* **2004**, *432*, 824–828.

[75] Reymond, J.-L.; Van Deursen, R.; Blum, L. C.; Ruddigkeit, L. Chemical space as a source for new drugs. *MedChemComm* **2010**, *1*, 30–38.

[76] Agrafiotis, D. K.; Rassokhin, D. N.; Lobanov, V. S. Multidimensional scaling and visualization of large molecular similarity tables. *Journal of Computational Chemistry* **2001**, *22*, 488–500.

[77] Peltason, L.; Iyer, P.; Bajorath, J. Rationalizing three-dimensional activity landscapes and the influence of molecular representations on landscape topology and the formation of activity cliffs. *Journal of chemical information and modeling* **2010**, *50*, 1021–1033.

[78] Peltason, L.; Bajorath, J. Systematic computational analysis of structure–activity relationships: concepts, challenges and recent advances. *Future medicinal chemistry* **2009**, *1*, 451–466.

[79] Kireeva, N.; Baskin, I.; Gaspar, H.; 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*, 301–312.

[80] Gaspar, H. A.; Marcou, G.; Horvath, D.; Arault, A.; Lozano, S.; Vayer, P.; Varnek, A. Generative topographic mapping-based classification models and their applicability domain: application to the biopharmaceutics drug disposition classification system (BDDCS). *Journal of chemical information and modeling* **2013**, *53*, 3318–3325.

[81] Horvath, D.; Brown, J.; Marcou, G.; Varnek, A. An evolutionary optimizer of libsvm models. *Challenges* **2014**, *5*, 450–472.

[82] Gaspar, H. A.; Baskin, I. I.; Marcou, G.; Horvath, D.; Varnek, A. GTM-Based QSAR Models and Their Applicability Domains. *Molecular Informatics* **2015**, *34*, 348–356.

[83] Horvath, D.; Koch, C.; Schneider, G.; Marcou, G.; Varnek, A. Local neighborhood behavior in a combinatorial library context. *Journal of computer-aided molecular design* **2011**, *25*, 237–252.

[84] Klimenko, K.; Marcou, G.; Horvath, D.; Varnek, A. Chemical Space Mapping and Structure–Activity Analysis of the ChEMBL Antiviral Compound Set. *Journal of Chemical Information and Modeling* **2016**, *56*, 1438–1454.

[85] Maggiora, G. M.; Bajorath, J. Chemical space networks: a powerful new paradigm for the description of chemical space. *Journal of computer-aided molecular design* **2014**, *28*, 795–802.

[86] Wawer, M.; Peltason, L.; Weskamp, N.; Teckentrup, A.; Bajorath, J. Structure- activity relationship anatomy by network-like similarity graphs and local structure- activity relationship indices. *Journal of medicinal chemistry* **2008**, *51*, 6075–6084.

[87] Peltason, L.; Bajorath, J. SAR index: quantifying the nature of structure-activity relationships. *Journal of medicinal chemistry* **2007**, *50*, 5571–5578.

[88] Peltason, L.; Hu, Y.; Bajorath, J. From structure–activity to structure–selectivity relationships: quantitative assessment, selectivity cliffs, and key compounds. *ChemMedChem* **2009**, *4*, 1864–1873.

[89] Dimova, D.; Wawer, M.; Wassermann, A. M.; Bajorath, J. Design of multitarget activity landscapes that capture hierarchical activity cliff distributions. *Journal of chemical information and modeling* **2011**, *51*, 258–266.

[90] Iyer, P.; Stumpfe, D.; Bajorath, J. Molecular mechanism-based network-like similarity graphs reveal relationships between different types of receptor ligands and structural changes that determine agonistic, inverse-agonistic, and antagonistic effects. *Journal of chemical information and modeling* **2011**, *51*, 1281–1286.

[91] Wawer, M.; Bajorath, J. Local structural changes, global data views: Graphical substructure- activity relationship trailing. *Journal of medicinal chemistry* **2011**, *54*, 2944–2951.

[92] Wawer, M.; Bajorath, J. Similarity- potency trees: a method to search for SAR information in compound data sets and derive SAR rules. *Journal of chemical information and modeling* **2010**, *50*, 1395–1409.

[93] Wollenhaupt, S.; Baumann, K. inSARa: Intuitive and interactive SAR interpretation by reduced graphs and hierarchical MCS-based network navigation. *Journal of chemical information and modeling* **2014**, *54*, 1578–1595.

[94] Gupta-Ostermann, D.; Hu, Y.; Bajorath, J. Introducing the LASSO graph for compound data set representation and structure–activity relationship analysis. *Journal of medicinal chemistry* **2012**, *55*, 5546–5553.

[95] Zhang, B.; Hu, Y.; Bajorath, J. AnalogExplorer: A New Method for Graphical Analysis of Analog Series and Associated Structure–Activity Relationship Information. *Journal of medicinal chemistry* **2014**, *57*, 9184–9194.

[96] Zwierzyna, M.; Vogt, M.; Maggiora, G. M.; Bajorath, J. Design and characterization of chemical space networks for different compound data sets. *Journal of computer-aided molecular design* **2015**, *29*, 113–125.

[97] Zhang, B.; Vogt, M.; Maggiora, G. M.; Bajorath, J. Comparison of bioactive chemical space networks generated using substructure-and fingerprint-based measures of molecular similarity. *Journal of computer-aided molecular design* **2015**, *29*, 595–608.

[98] Zhang, B.; Vogt, M.; Maggiora, G. M.; Bajorath, J. Design of chemical space networks using a Tanimoto similarity variant based upon maximum common substructures. *Journal of computer-aided molecular design* **2015**, *29*, 937–950.

[99] Kunimoto, R.; Vogt, M.; Bajorath, J. Maximum common substructure-based Tversky index: an asymmetric hybrid similarity measure. *Journal of computer-aided molecular design* **2016**, *30*, 523–531.

[100] de la Vega de León, A.; Bajorath, J. Chemical space visualization: transforming multidimensional chemical spaces into similarity-based molecular networks. *Future Medicinal Chemistry* **2016**, *8*, 1769–1778.

[101] Wassermann, A. M.; Haebel, P.; Weskamp, N.; Bajorath, J. SAR matrices: automated extraction of information-rich SAR tables from large compound data sets. *Journal of chemical information and modeling* **2012**, *52*, 1769–1776.

[102] Wermuth, C. G. *The practice of medicinal chemistry*; Academic Press, 2011.

[103] Stumpfe, D.; Bajorath, J. Methods for SAR visualization. *RSC Advances* **2012**, *2*, 369–378.

[104] Stumpfe, D.; Bajorath, J. Recent developments in SAR visualization. *MedChemComm* **2016**, *7*, 1045–1055.

[105] Maynard, A. T.; Roberts, C. D. Quantifying, Visualizing, and Monitoring Lead Optimization. *Journal of medicinal chemistry* **2015**, *59*, 4189–4201.

[106] Reutlinger, M.; Guba, W.; Martin, R. E.; Alanine, A. I.; Hoffmann, T.; Klenner, A.; Hiss, J. A.; Schneider, P.; Schneider, G. Neighborhood-Preserving Visualization of Adaptive Structure–Activity Landscapes: Application to Drug Discovery. *Angewandte Chemie International Edition* **2011**, *50*, 11633–11636.

[107] Medina-Franco, J. L.; Mart\'\inez-Mayorga, K.; Bender, A.; Mar\'\in, R. M.; Giulianotti, M. A.; Pinilla, C.; Houghten, R. A. Characterization of activity landscapes using 2D and 3D similarity methods: consensus activity cliffs. *Journal of chemical information and modeling* **2009**, *49*, 477–491.

[108] Maggiora, G. M. On outliers and activity cliffs why QSAR often disappoints. 2006.

[109] Vogt, M.; Stumpfe, D.; Maggiora, G. M.; Bajorath, J. Lessons learned from the design of chemical space networks and opportunities for new applications. *Journal of computer-aided molecular design* **2016**, *30*, 191–208.

[110] Griffen, E.; Leach, A. G.; Robb, G. R.; Warner, D. J. Matched Molecular Pairs as a Medicinal Chemistry Tool: Miniperspective. *Journal of medicinal chemistry* **2011**, *54*, 7739–7750.

[111] Liu, T.; Lin, Y.; Wen, X.; Jorissen, R. N.; Gilson, M. K. BindingDB: a web-accessible database of experimentally determined protein–ligand binding affinities. *Nucleic acids research* **2007**, *35*, D198–D201.

[112] Gamo, F.-J.; Sanz, L. M.; Vidal, J.; de Cozar, C.; Alvarez, E.; Lavandera, J.-L.; Vanderwall, D. E.; Green, D. V. S.; Kumar, V.; Hasan, S.; Others, Thousands of chemical starting points for antimalarial lead identification. *Nature* **2010**, *465*, 305–310.

[113] OEChem, T. K. OpenEye Scientific Software. *Inc., Santa Fe, NM, USA* **2012**,

[114] Oprisiu, I.; Varlamova, E.; Muratov, E.; Artemenko, A.; Marcou, G.; Polishchuk, P.; Kuz'Min, V.; Varnek, A. QSPR approach to predict nonadditive properties of mixtures. application to bubble point temperatures of binary mixtures of liquids. *Molecular Informatics* **2012**, *31*, 491–502.

[115] O'Madadhain, J.; Fisher, D.; Smyth, P.; White, S.; Boey, Y.-B. Analysis and visualization of network data using JUNG. *Journal of Statistical Software* **2005**, *10*, 1–35.

[116] Fruchterman, T. M. J.; Reingold, E. M. Graph drawing by force-directed placement. *Software: Practice and experience* **1991**, *21*, 1129–1164.

[117] Heikamp, K.; Hu, X.; Yan, A.; Bajorath, J. Prediction of activity cliffs using support vector machines. *Journal of chemical information and modeling* **2012**, *52*, 2354–2365.

[118] ChemAxon: Budapest, H. ChemAxon API.

[119] Strasbourg Virtual Screening Web Server. `http://infochim.u-strasbg.fr/webserv/VSEngine.html`, Accessed: 01-07-2016.

Declaration

I herewith certify that this material is my own work, that I used only those sources and resources referred to in the thesis, and that I have identified citations as such.

October 9, 2017

Shilva Kayastha