J Comput Aided Mol Des (2014) 28:597–618
DOI 10.1007/s10822-014-9743-1

Computational chemogenomics: Is it more than inductive transfer?

J. B. Brown · Yasushi Okuno · Gilles Marcou · Alexandre Varnek · Dragos Horvath

Received: 20 January 2014 / Accepted: 11 April 2014 / Published online: 27 April 2014

Abstract High-throughput assays challenge us to extract knowledge from multi-ligand, multi-target activity data. In QSAR, weights are statically fitted to each ligand descriptor with respect to a single endpoint or target. However, computational chemogenomics (CG) has demonstrated benefits of learning from entire grids of data at once, rather than building target-specific QSARs. A possible reason for this is the emergence of inductive knowledge transfer (IT) between targets, providing statistical robustness to the model, with no assumption about the structure of the targets. Relevant protein descriptors in CG should allow one to learn how to dynamically adjust ligand attribute weights with respect to protein structure. Hence, models built through explicit learning (EL) by including protein information, while benefitting from IT enhancement, should provide additional predictive capability, notably for protein deorphanization. This interplay between IT and EL in CG modeling is not sufficiently studied. While IT is likely to occur irrespective of the injected target information, it is not clear whether and when boosting due to EL may occur. EL is only possible if protein description is appropriate to the target set under investigation. The key issue here is the search for evidence of genuine EL exceeding expectations based on pure IT. We explore the problem in the context of Support Vector Regression, using more than 9,400 pK_i values of 31 GPCRs, where compound–protein interactions are represented by the concatenation of vectorial descriptions of compounds and proteins. This provides a unified framework to generate both IT-enhanced and potentially EL-enabled models, where the difference is toggled by supplied protein information. For EL-enabled models, protein information includes genuine protein descriptors such as typical sequence-based terms, but also the experimentally determined affinity cross-correlation fingerprints. These latter benchmark the expected behavior of a quasi-ideal descriptor capturing the actual functional protein-protein relatedness, and therefore thought to be the most likely to enable EL. EL- and IT-based methods were benchmarked alongside classical QSAR, with respect to cross-validation and deorphanization challenges. A rational method for projecting benchmarked methodologies into a strategy space is given, in the aims that the projection will provide directions for the types of molecule designs possible using a given methodology. While EL-enabled strategies outperform classical QSARs and favorably compare to similar published results, they are, in all respects evaluated herein, *not* strongly distinguished from IT-enhanced models. Moreover, EL-enabled strategies failed to prove superior in deorphanization challenges. Therefore, this paper raises caution that, contrary to common belief and intuitive expectation, the benefits of chemogenomics models over classical QSAR are quite possibly due less to the injection of protein-related information, and rather impacted more by the effect of inductive transfer, due to simultaneous learning from all of the modeled endpoints. These results show that the field of protein descriptor research needs further improvements to truly realize the expected benefit of EL.

Electronic supplementary material The online version of this article (doi:10.1007/s10822-014-9743-1) contains supplementary material, which is available to authorized users.

J. B. Brown · Y. Okuno
Department of Clinical System Onco-Informatics, Graduate School of Medicine, Kyoto University, Kyoto 606-8501, Japan

G. Marcou · A. Varnek · D. Horvath (✉)
Laboratoire de Chémoinformatique, UMR 7140 CNRS, Univ. Strasbourg, 1, rue Blaise Pascal, 67000 Strasbourg, France
e-mail: d.horvath@unistra.fr

Keywords Chemogenomics · Proteochemometrics · QSAR · Machine learning · Inductive transfer

Abbreviations

CG	Chemogenomics
GA	Genetic algorithm
DS	Descriptor space
GPCR	G-protein coupled receptor
QSAR	Quantitative structure–activity relationships
SVM	Support vector machine
SVR	Support vector regression
IT	Inductive transfer
MTL	Multi-task learning
RMSE	Root mean squared error
ISIDA	In silico design and data analysis

Introduction

Advances in high-throughput technologies have enabled us to generate enormous volumes of cellular, functional, and target-specific bioactivity data for compounds. Despite these advances, there is still a considerable gap between our ability to link this data as a whole to phenotypical outcomes, particularly with respect to unintended drug side effects. It also frequently occurs that experimental researchers execute assays on both cellular and target-specific levels using a fixed set of compounds, but are left to only speculate about the connection between the two outcomes. In these situations, development of models that explain patterns or correlation between the two levels of experimental data can be beneficial.

Computational `chemogenomics (CG)` or `proteochemometrics` (terms considered synonymous in this work) recently emerged as the paradigm of learning from polypharmacological (multi-target) profiles [7, 8, 30, 51]. It characterizes each putative ligand–target complex by a composite set of both small-molecule descriptors encoding the ligand and protein descriptors encoding the target [24, 26]. As increasingly more emphasis is set on understanding and early prediction of drug side effects [42, 43], CG naturally emerged as an attempt to address such questions, spurred by steadily accumulating multi-target activity profile data due to routine screening of pre-drug candidates over a wealth of potentially relevant biological targets.

It is important to mention that in CG (seen as the QSAR of protein–ligand complexes) both ligand and target may, formally, be considered as equivalent. CG may serve both to predict ligand affinity in virtual screening [45] of a compound collection against a given target, as well as predict protein affinity in attempts to find novel targets that may bind a given drug (drug repositioning [19]).

Various approaches to encode protein sequence and structure under the form of numeric descriptors have been suggested. While 3D structure-based descriptors, exploiting knowledge about the location and the geometry of the binding site [40, 48], are clearly the most information-rich, they are practically not very useful for target deorphanization (targets with well-characterized binding sites cannot qualify as "orphans"). An interesting alternative is represented by the injection of information of known or assumed key binding site residues into the protein fingerprint [13, 47]. The most general approach, not making any a priori assumption about the targets, include empirical amino acid sequence-derived descriptors, also used in protein sequence–activity relationship modeling [11, 37].

In the CG formalism, target and ligand descriptors are perfectly interchangeable. However, in this paper we will nevertheless adopt, a ligand-centric approach to CG for three main reasons. First, virtual screening for novel ligands is more often employed than the orthogonal search for new targets of a ligand. Next, classical ligand-centric QSAR is actively employed in the herein reported benchmarking study. Last but not least, the provided protein information does not have to be structurally relevant in order to benefit from multi-target learning, as will be emphasized in the following. Therefore, this work will describe CG as a ligand-centric approach, where the ligand structure–activity relationships are allegedly "modulated" by protein information. This point of view serves for discussion only, and has no impact on the computational strategies and their results.

The naive approach to multi-target profile prediction would simply consist in realizing, for each target T, an individual QSAR model $\hat{A}_T(M) = f[\mathbf{D}(M)]$ (circumflex cap meaning in silico calculated value throughout this text), where a molecule M is described by $\mathbf{D}(M)$. In their simplest form, such relationships are linear:

$$\hat{A}(M) = \alpha_0 + \alpha_1 D_1(M) + \alpha_2 D_2(M) + \cdots + \alpha_n D_n(M) \tag{1}$$

where coefficients (weights) α_i represent the relative impacts of the ligand feature i encoded by D_i of the activity A on the current target. They will be termed "feature weights" in the following.

Although this work exclusively deals with non-linear models, the linear approach above will be used to illustrate the concepts that are central to this work. These concepts are independent of the actual functional form of the models, and hence easiest explained on the basis of maximum simplicity approaches. The reader is encouraged to visit the data mining literature for a more formal treatment [1, 2, 9].

Predicting activity $\hat{A}$ of ligand M by individual models trained for each T will mechanically allow the completion of the ligand-by-target matrix of predicted activities. This amounts to determining the matrix of feature weights for targets (symbolically, α_i^t), by successively fitting each vector $\boldsymbol{\alpha}^t$ for every t. Here, each α_i^t is *implicitly* associated to a target t, the unique data source serving to fit its value. No explicit knowledge on the relationships between these weights and the nature of the target is generated here. Fitting (α_i^t) assumes that the initial training data supports—in terms of the per-protein ligand set size and diversity (chemical space coverage)—the building of all of these individual models. Unfortunately, experimental activity profiles $A_T(M)$ are often sparse; few M, T pairs were subjected to experimental scrutiny, and fewer still provide examples of high-affinity complexes. Moreover, the only way to update the α_i^t matrix in order to cover a novel target T is by fitting $\boldsymbol{\alpha}^T$ values. This is conditioned by the existence of sufficient and diverse training examples of binders and non-binders, thus obviously unfeasible for orphan targets.

However, the paradigm of CG, that is the simultaneous machine learning from the entire available activity matrix, may significantly outperform the above-mentioned naive strategy. There are two main reasons for this:

1. *Inductive transfer* (*IT*) Under its most simple form, the principle of IT [44, 50] can be outlined as follows. Suppose that the activity of ligands with respect to a target t is conditioned by n ligand features, as highlighted in Eq. 1. Now consider a related target T, depending on exactly the same n features, where selectivity stems from a single feature that is weighted differently: $\hat{A}_T(M) = \beta_1 D_1(M) + \alpha_2 D_2(M) + \cdots$. In the naive approach, fitting either equation would require, by rule-of-thumb, $20n$ or more training examples in order to grant some statistical robustness to the models [16]. The strength of IT is the transfer of knowledge (e.g., α_i values) obtained from analysis of a related problem to enhance solving of a new one. Suppose that enough data is available to train $\hat{A}_t$. Then, predicted values $\hat{A}_t$ could be employed as a new molecular descriptor for the related model $\hat{A}_T = \hat{A}_t + (\beta_1 - \alpha_1)D_1$. A few ligands, including several actives, tested on T would suffice to robustly determine the two coefficients (1.0 and $\beta_1 - \alpha_1$) of the latter regression.

   In practice, IT is not bound to such a sequential training scheme as outlined above (output of primary models serving as input for the IT-enhanced approaches, often referred to as "feature nets" [44]), but may also be achieved by simultaneous (multi-task [41]) learning for related sets of tasks (endpoints). Related tasks share a common latent subspace of descriptors, enhanced with elements responsible for the specificity of each target.
2. *Explicit learning* (*EL*) This requires target structural information to be injected into the learning process, by means of a protein descriptor vector $\boldsymbol{\Delta}(T)$. Intuitively, one may think about an EL model as a QSAR equation in which the weights $\alpha_i^t = f_i[\boldsymbol{\Delta}(t)]$ are now functions of the protein descriptors. The reasoning for this is that the relation between ligand structures and their affinities to observable endpoints should be related to and explained by the explicit provision of the protein information. As each protein responds differently to the presence of a ligand feature i (a substructure, for example), the intensity of this response is dependent on protein structure. If the relevant protein structure features are captured by the descriptor $\boldsymbol{\Delta}$, then the conditioning of α_i with respect to $\boldsymbol{\Delta}$ should be learned during CG model building, leading to true EL-enhanced models. For example [23, 25], if affinity is represented as a linear combination of ligand and protein descriptor cross-terms, as shown in Eq. (2):

$$\hat{A}(m,t) = \alpha_0 + \sum_{i,j} \beta_{ij} \times D_i(m)\Delta_j(t) \tag{2}$$

   then $\alpha_i^t = \sum_j \beta_{ij}\Delta_j(t)$.

Both the IT and EL concepts are independent of machine learning approaches, and therefore should be insertable into any given algorithm, including, but not limited to, the well-known support vector machines (SVMs) [12, 21, 22, 34] or neural networks [41, 44]. Hence, it is possible to employ the same algorithm to generate both IT-enhanced and potentially EL-enabled models, where the difference is toggled by supplied protein information. The key difference between the IT-enhanced and EL-enabled approaches is that the former, like naive single-endpoint QSARs, are completely ignorant of the nature of the targets. The IT approaches are injected with indicator variables which replace actual protein information. In other words, $\boldsymbol{\Delta}(t)$ should stand for actual physicochemical and structural target properties in EL, while proteins are represented by mere labels in IT. For example, in previous work led by Vert [21, 22], EL-enabled models rely on calculated or biology-inspired protein kernel values [27, 38], while IT-enhanced models are based on the so-called "multi-task" kernel. These studies are, to our knowledge, the most extensive analysis of IT versus EL, done in terms of classification, based on the "kernel trick". This "trick" operates under the tensor-product working hypothesis: if ligand–target complexes denoted by $m : t$ are described by the tensor $\mathbf{D}(m) \otimes \mathbf{\Delta}(t)$, then the expensive cross-product calculation can be avoided by alternatively computing the product of ligand and target kernels:

$$\mathcal{K}(m:t, M:T) = \mathcal{K}^{lig}(m, M) \times \mathcal{K}^{prot}(t, T) \quad (3)$$

Despite reported success in previous studies, it is yet to be made clear whether injection of actual protein information in intended EL models actually leads to the desired EL model, or whether machine learning merely exploits those protein descriptors in the same way as it would handle target labels in IT processes. Moreover, the few realizations [23, 25] of explicit EL-enhanced approaches based on Eq. (2) only marginally outperformed the simpler linear combination of stand-alone ligand and protein descriptors. This is further evidence that granting the technical feasibility for the fitting of an EL-enhanced approach is not a guarantee that the resulting model will actually attain such status. The present work aims to shed some more light on this issue.

To this purpose, a rigorous benchmarking protocol was designed, in order to compare (a) single-endpoint QSAR models, (b) IT-enhanced single-endpoint QSARs, (c) IT-enhanced (multi-endpoint) CG models, and eventually (d) EL-enabled (protein information-supplied) CG models. This should enable us to weight multiple perspectives for discovering hidden knowledge in phenotypical and other endpoint assay data.

We used this opportunity to focus on quantitative support vector regression (SVR) [20, 39]. This is more challenging than SVM classification, and such studies were so far, only performed as proof-of-concept [3, 4, 28]. To our knowledge, SVR was never explicitly investigated in the context of IT versus EL benchmarks. Like in the above-mentioned SVR-driven studies, we opted for the maximum simplicity option for ligand–target descriptor pairs: concatenation of their respective descriptors. This means that a putative ligand–protein complex $M : T$ will be considered as one object represented by a vector $\mathbf{D}(M), \mathbf{\Delta}(T)$ resulting from concatenation of ligand and protein terms respectively.

Based on 9,642 accurate GPCR-ligand complexes of measured pK_i values (approximately 4,500 ligands for 31 rhodopsin-like GCPRs), the herein used training set is one of the largest coherent multi-target sets seen in CG studies so far. Featuring actual high and low-affinity ligand-protein pairs, it has no need to rely on artificially generated, experimentally untested entries as decoys.

ISIDA property-labeled fragment counts [36] and fuzzy pharmacophore triplets [5, 6] were used to describe ligands. In terms of genuine protein descriptors, we employed a two-pronged approach. On one hand, we utilize sequence-based terms that can be easily calculated for poorly-studied proteins, and are thus potentially usable in a non-simulated, true deorphanization attempt. On the other hand, a fingerprint of protein–protein affinity-focused similarity scores based on directly measured experimental affinities was exploited. EL-enabled models in our approach use the genuine protein descriptors, whereas IT-enhanced models introduce indicator variables (identity fingerprints); both are concatenated to ligand descriptors.

The analyses executed were as follows. First, the various approaches have been benchmarked in terms of model building and cross-validation propensities. A challenging SVR cross-validation protocol was based on a 10-trial randomized leave-1/3-out scheme. A genetic algorithm (GA)-driven optimization of the SVR operational parameters has been employed to build optimal models within each of the CG modeling approaches tested. Cross-validated prediction propensities of models were monitored in terms of residual errors, allowing us to locate benchmarked machine learning strategies in a "strategy space" and to report mutual closeness relationships.

Second, a target deorphanization study was carried out. Whilst there is not a priori expectation to see EL-enabled models outperform IT-enhanced approaches in terms of cross-validation propensities, target deorphanization challenges are the actual stumbling block for genuine EL-enabled CG. Indeed, EL should display a decisive advantage, for it is expected to explicitly adapt the $\alpha_i^t = f_i[\mathbf{\Delta}(t)]$ weights to the orphan target. By contrast, as already mentioned, IT methods are unaware of the nature of orphans. However, while a rigorous deorphanization protocol cannot proceed without providing relevant target information, some less rigorous alternatives do exist. A baseline deorphanization strategy, herein termed "deorphanization by substitution", advocates using a predictive model for some training set proteins as a predictor of the affinity of the presumedly orphan protein. While there is no fundamental reason for the success of such a strategy, in practice this may well be the case, if the presumed orphan has at least some close analogs among training set proteins (c.f., [46]). Unfortunately, published deorphanization success stories rarely explicitly report the closeness to training set proteins, or how well the single-endpoint models of those proteins would have fared instead of the advocated CG approach [21, 22]. Therefore, the target deorphanization protocol here assigns paramount importance to this aspect.

Our study has interestingly revealed that, while CG methods once more confirmed their advantages over classical QSAR, most of this advantage seems to be due to IT. No significant boost of EL approaches over IT strategies was evidenced. The direct consequence of this is that successful deorphanization was restricted to ‘trivial’ cases of targets being quite close to one or several training set proteins. It further shows that future CG studies should benchmark EL against some baseline IT experiment to provide sufficient proof of EL, and that the field of protein descriptor research needs further improvements to truly realize the expected benefit of EL. A future implication of this is that once one has established proof of concept for an EL-enabled model, they can return to the critical overarching task of applying their CG model to molecule design.

Methods

This section begins with the presentation of the data—ligands and targets—followed by a brief overview of respective ligand and protein descriptors. Next, a general description of model building is given, prior to the actual introduction of the various computational strategies used. The comparison criteria used in benchmarking of these approaches will be then reported, before outlining the last key topic, the target deorphanization protocol.

The chemogenomics profile: targets, ligands, data

This study concerns $N_T = 31$ different rhodopsin-like G-protein coupled receptors (GPCRs), listed in Table 1 together with their ChEMBL [14] codes. They were selected due to the relative wealth of pK_i affinity constants (measured for at least 50, and up to more than 1,000 ligands) in ChEMBL. For each target, the associated sets of ligands of known pK_i were graciously provided by Prof. Jürgen Bajorath, University of Bonn, after curation. In total, a set of 9,462 $pK_i(m, T)$ values reporting the affinity of ligands m with respect to targets T were available. Given that some targets have overlapping ligands, and that stereochemistry was not accounted for by molecular descriptors used, the number of unique ligands was 4,685.

Table 1 List of herein employed GPCR targets; N_{lig} is the number of ligands available for a given target

T	Target name	ChEMBL ID	N_{lig}
1	Alpha-1a adrenergic receptor	229	222
2	Alpha-1b adrenergic receptor	232	141
3	Alpha-1d adrenergic receptor	223	130
4	Alpha-2a adrenergic receptor	1867	158
5	Alpha-2b adrenergic receptor	1942	94
6	Alpha-2c adrenergic receptor	1916	119
7	Beta-1 adrenergic receptor	213	168
8	Beta-2 adrenergic receptor	210	221
9	Beta-3 adrenergic receptor	246	141
10	Dopamine D1 receptor	2056	272
11	Dopamine D2 receptor	217	1,325
12	Dopamine D3 receptor	234	846
13	Dopamine D4 receptor	219	424
14	Dopamine D5 receptor	1850	98
15	Melatonin receptor 1A	1,945	215
16	Melatonin receptor 1B	1946	262
17	Muscarinic acetylcholine receptor M1	216	151
18	Muscarinic acetylcholine receptor M2	211	188
19	Muscarinic acetylcholine receptor M3	245	319
20	Muscarinic acetylcholine receptor M4	1821	67
21	Muscarinic acetylcholine receptor M5	2035	51
22	Serotonin 1a (5-HT1a) receptor	214	884
23	Serotonin 1b (5-HT1b) receptor	1898	138
24	Serotonin 1d (5-HT1d) receptor	1983	139
25	Serotonin 2a (5-HT2a) receptor	224	654
26	Serotonin 2b (5-HT2b) receptor	1833	256
27	Serotonin 2c (5-HT2c) receptor	225	504
28	Serotonin 4 (5-HT4) receptor	1875	62
29	Serotonin 5a (5-HT5a) receptor	3426	79
30	Serotonin 6 (5-HT6) receptor	3371	859
31	Serotonin 7 (5-HT7) receptor	3155	275

The number of unique ligands in the dataset is 4,685

Ligand descriptors

Several different descriptor spaces (DS) $\mathbf{D}(M)$ were considered to represent ligands.

1. Fuzzy pharmacophore triplets (FPT) [5, 6] represent fuzzy counts of monitored triplets of pharmacophore features (hydrophobe, aromatic, H-bond donor and acceptor, cation and anion), at given topological inter-feature distances. Out of the considered FPT setups discussed in the original publication [6], the default FPT1 was employed here.
2. Property-labeled ISIDA fragment descriptors were introduced [36] as a generalization of molecular substructure counts. Fragment counts were named according to the initial convention, namely ‘fragment type ff + coloring property PP + lower and upper fragment size lu’. The fragment type may be one of:

   `seq`—sequence
   `aa`—augmented atom i.e. atom-centered, circular fragments)
   `tree`—trees, handled in this work augmented atoms in which the nature of all atoms except the center and the terminal ‘leaves’ are ignored.

   The presence of a ‘b’ label following the fragment types means that bond order information is taken into account (ignored by default). Next, ISIDA coloring properties may be:

`SY` atomic symbols
`PH` pH-sensitive pharmacophore type at pH 7.4

Finally, the following two digits refer to the lowest and the highest sequence length or circular fragment radius. As a complete example, `seqSY37` are counts of symbol-colored sequences of 3–7 atoms, such as ‘CCNCC’.

Note that all pH-sensitive descriptor sets are, as shown in cited works, not simple colored subgraph or triplet counts, but fuzzy average estimators of the overall occupancy of corresponding subgraphs/triplets in a population of the various microspecies of the molecule at proteolytic equilibrium in a neutral buffer of pH 7.4.

The six descriptor sets effectively used in this work are: FPT1, seqPH37, seqSY37, aabPH02, treePH03, treeSY03. For more details on these particular terms, please refer to previous publications in which they were employed [18].

Target descriptors and fingerprints

Target spaces $\mathbf{\Delta}(T)$ include both regular protein descriptors for EL-enabled models and identity fingerprints for labeling each pK_i entry by its associated target in IT-enhanced CG.

The herein employed protein descriptors are as follows.

1. `PROFEAT` [29, 33] are sequence-based protein descriptors including amino acid composition, dipeptide composition, normalized Moreau–Broto autocorrelation, Moran autocorrelation, Geary autocorrelation, sequence-order-coupling number, quasi-sequence-order descriptors and the composition, transition and distribution of various structural, and physicochemical properties. They were generated by submitting the ChEMBL-retrieved sequences of the 31 targets to the PROFEAT descriptor web server. This resulted in a 1,495-dimensional PROFEAT descriptor vector, after exclusion of elements of null variance.
2. `seqAA` amino acid sequence counts were ad-hoc generated by a text analysis script monitoring all of the two- to six-letter substrings of the one-letter-code protein sequence strings. Each of the 8,485 herewith generated descriptor elements represents the number of non-zero occurrences of the associated di- to hexapeptide in a protein.
3. The `SIMFP` target similarity fingerprint is the real-valued $N_T = 31$-dimensional vector functional descriptor having as element t a relatedness score of the current target T to protein t of the profile. $SIMFP_T(t)|_{t\neq T}$ is defined as the Pearson correlation coefficient of respective experimental $pK_i(M,T)$ and $pK_i(M,t)$ over the subset of common binders M, if more than 15 common binders exist in data set for T and t. Otherwise, the value is set to zero.
4. The `IDFP` identity fingerprint is a simple N_T-dimensional binary vector in which only the bit associated to the current target is highlighted; formally $IDFP_T(t) := (T == t)$, which is also known as the Kronecker delta function. It is a pure indicator variable, carrying no other information but the target ID. In the IDFP target space, each protein is identical to itself, and maximally distant to all the others.

The first three entries provide genuine protein information, and thus serve in EL-enabled models. By contrast, IDFP-based CG models do not profit from other sources of enhancement but only from IT.

Model building

All machine learning was based on the widely-used `libsvm` software [10], in *epsilon*-regression mode, with control parameters optimized by a GA [32], in order to maximize the cross-validation robustness of the obtained models. A set of control parameters is hereafter referred to as a “chromosome”. Realized models differ only in terms of the descriptors or descriptor combinations submitted for learning, but follow strictly the same 10-trial randomized leave-1/3-out cross-validation procedure.

The first locus of the chromosome encodes the ligand DS to be used. Several DS may be allowed to compete against each other for the status of “best model provider”. The second locus is a scaling toggle, defining whether these descriptors should be used as such, or subjected to a min/max scaling, using the libsvm `svm-scale` tool. The following loci encode for specific parameters of SVR: `kernel` [linear, polynomial, radial basis function (RBF) or Sigmoid] `cost` value (real value, allowed to vary between 10^{-9} and 10^{9}), `gamma` and `coef0`. The `gamma` value needs to be chosen with respect to the average distance values between the training set points in the current DS. Therefore, average Euclidean and dot product distances are calculated and the actual gamma passed on to SVR will equal the value contained by the chromosome locus (which may range between 0.2 and 6.0) divided by the suited average value (Euclidean when the Gaussian kernel is used, dot product otherwise). The coef0 offset (relevant for polynomial and sigmoidal kernels only) ranges between -10 and 10. Epsilon, controlling the loss function in SVR, was kept to the default 0.1, which represents 10 % of one log of standard deviation of training pK_i values.

At a given operational setup as encoded in the current chromosome, cross-validated (XV) machine learning is performed. This leads to building a total of 10 × 3 individual models. At the end, the procedure returns, for each item, 10 "external" and 20 "fitted" estimations.

In order to estimate the robustness of the model associated to the given chromosome, for each of the $r = \{1\ldots10\}$ groups of external prediction values, root-mean-squared errors $(RMSE_r)$ and associated determinations coefficients Q_r^2 are calculated The respective average and standard deviation $\langle Q_r^2 \rangle$ and $\sigma(Q_r^2)$ are taken over the $r = 10$ XV repeats, and the fitness of the chromosome is asses by

$$\Theta^2 = \langle Q_r^2 \rangle - 2 \times \sigma(Q_r^2) \tag{4}$$

where the determination coefficient Q_r^2 for experiment r is taken as the complement of the ratio of the squared $RMSE_r$ versus the variance of the pKi values of the data set. This fitness function (4) is designed to favor high-accuracy, high-repeatability strategies. The resulting model of best Θ^2 is thus a consensus approach based on the 30 above-mentioned individual models. Using the model for external prediction of an item outside of the training set amounts to applying each of the individual models, taking the average of the 30 estimated activity values as predicted output $\hat{A}$.

Consensus and unbiased predictors

When an item from the training set of $\hat{A}$ is submitted for prediction by this model as part of some external set, the returned consensus will be dominated by fitted values. For the sake of a more stringent evaluation of predictive power, fitted values should be ignored in benchmarking studies. In contrast to the consensus property estimator $\hat{A}$, we will define the unbiased estimator $\tilde{A}$ as a model returning $\hat{A}$ for any item not part of any training set, but only the average of 10 external values otherwise. Model comparisons will be exclusively be carried out in terms of $\tilde{A}$ predicted values.

Learning strategies

This section describes the different computational experiments performed in this work, an overview of which is given in Table 2. A schematic depiction of the procedures, regrouped by categories, is organized as follows: ligand-only "classical" QSAR approaches (Fig. 1), IT-enhanced methods (Fig. 2) and EL-enabled approaches (Fig. 3).

1. `BQSAR`—Best single-endpoint QSAR models were generated (Fig. 1a) for each target T, allowing the algorithm to select the best descriptor space adapted to model each target.
2. `QSAR`—Best QSAR models (Fig. 1b) in the consensus DS `treePH03`.
3. `FQ`—Family QSAR (Fig. 1c) predicts the generic likelihood [15] of a ligand to bind to GPCRs, without specifically targeting any specific receptor. It was trained on the fused set of all ligands over all targets, with each ligand listed every time it appears in a target subset, associated to the pK_i of that target. No protein information is present.
4. `SE-IT`—Strong explicit inductive transfer models (Fig. 2a) are single-endpoint QSAR models including, for each T the BQSAR-predicted pK_i values of the ligand for other targets t, as additional descriptors. Targets t exclude, obviously, the current T, as well as those for which no relevant BQSAR model could be built. Out of $N_T = 31$ receptors, 28 returned models of

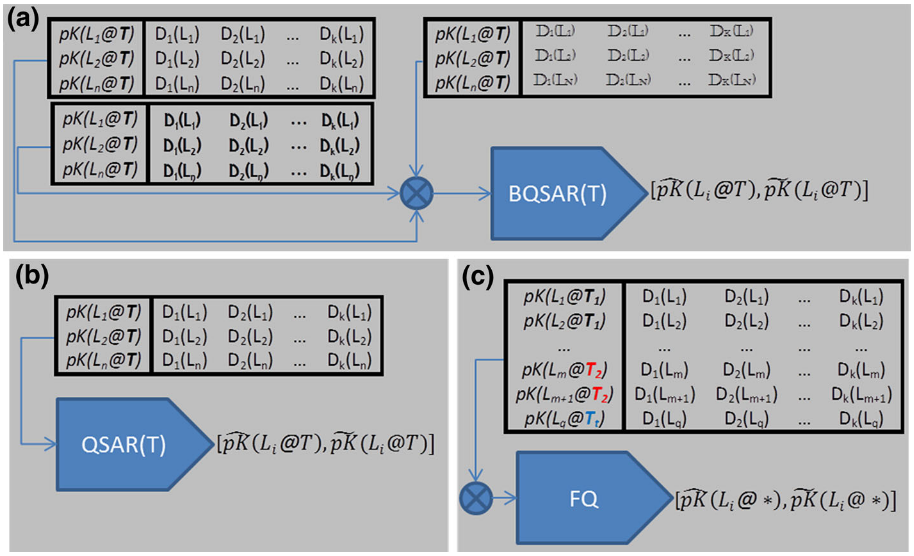
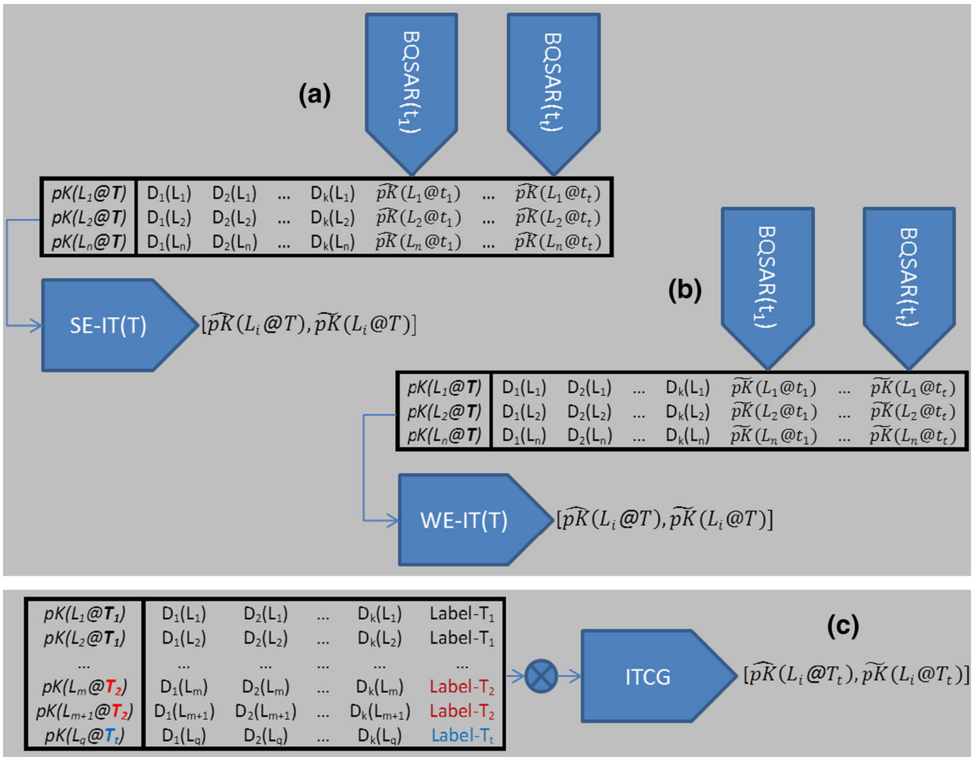

(a)
pK(L1@T) D1(L1) D2(L1) ... Dk(L1)
pK(L2@T) D1(L2) D2(L2) ... Dk(L2)
pK(Ln@T) D1(Ln) D2(Ln) ... Dk(Ln)
pK(L1@T) D1(L1) D2(L1) ... Dk(L1)
pK(L2@T) D1(L2) D2(L2) ... Dk(L2)
pK(Ln@T) D1(Ln) D2(Ln) ... Dk(Ln)
pK(L1@T) D1(L1) D2(L1) ... DK(L1)
pK(L2@T) D1(L2) D2(L2) ... DK(L2)
pK(Ln@T) D1(LN) D2(LN) ... DK(LN)
BQSAR(T)
$[\widehat{pK}(L_i@T), \widetilde{pK}(L_i@T)]$
(b)
pK(L1@T) D1(L1) D2(L1) ... Dk(L1)
pK(L2@T) D1(L2) D2(L2) ... Dk(L2)
pK(Ln@T) D1(Ln) D2(Ln) ... Dk(Ln)
QSAR(T)
$[\widehat{pK}(L_i@T), \widetilde{pK}(L_i@T)]$
(c)
pK(L1@T1) D1(L1) D2(L1) ... Dk(L1)
pK(L2@T1) D1(L2) D2(L2) ... Dk(L2)
... ...
pK(Lm@T2) D1(Lm) D2(Lm) ... Dk(Lm)
pK(Lm+1@T2) D1(Lm+1) D2(Lm+1) ... Dk(Lm+1)
pK(Lq@Tt) D1(Lq) D2(Lq) ... Dk(Lq)
FQ
$[\widehat{pK}(L_i@*), \widetilde{pK}(L_i@*)]$

Fig. 1 Schemes of herein employed "classical" ligand-only QSAR schemes. The *circled times* symbol signals that DS choice is enabled at model building step, even though explicit depiction of various candidate DS, with D vectors rendered in various typesets, was only provided for BQSAR. Any model may be used to return consensus (*hat operator*) or cross-validated (*tilde operator*) predictions

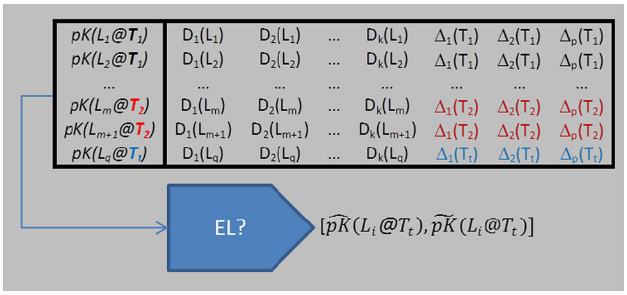

(a)
BQSAR(t_1)
BQSAR(t_t)
SE-IT(T)
$[\widehat{pK}(L_i@T),\widehat{pK}(L_i@T)]$
(b)
BQSAR(t_1)
BQSAR(t_t)
WE-IT(T)
$[\widetilde{pK}(L_i@T),\widetilde{pK}(L_i@T)]$
(c)
ITCG
$[\widehat{pK}(L_i@T_t),\widetilde{pK}(L_i@T_t)]$

$pK(L_1@T)$	$D_1(L_1)$	$D_2(L_1)$	...	$D_k(L_1)$	$\widehat{pK}(L_1@t_1)$	...	$\widehat{pK}(L_1@t_t)$
$pK(L_2@T)$	$D_1(L_2)$	$D_2(L_2)$	...	$D_k(L_2)$	$\widehat{pK}(L_2@t_1)$	...	$\widehat{pK}(L_2@t_t)$
$pK(L_n@T)$	$D_1(L_n)$	$D_2(L_n)$	...	$D_k(L_n)$	$\widehat{pK}(L_n@t_1)$	...	$\widehat{pK}(L_n@t_t)$

$pK(L_1@T)$	$D_1(L_1)$	$D_2(L_1)$	...	$D_k(L_1)$	$\widetilde{pK}(L_1@t_1)$	...	$\widetilde{pK}(L_1@t_t)$
$pK(L_2@T)$	$D_1(L_2)$	$D_2(L_2)$	...	$D_k(L_2)$	$\widetilde{pK}(L_2@t_1)$	...	$\widetilde{pK}(L_2@t_t)$
$pK(L_n@T)$	$D_1(L_n)$	$D_2(L_n)$	...	$D_k(L_n)$	$\widetilde{pK}(L_n@t_1)$	...	$\widetilde{pK}(L_n@t_t)$

$pK(L_1@T_1)$	$D_1(L_1)$	$D_2(L_1)$	...	$D_k(L_1)$	Label-T_1
$pK(L_2@T_1)$	$D_1(L_2)$	$D_2(L_2)$	...	$D_k(L_2)$	Label-T_1
...	...	...	...	...	...
$pK(L_m@T_2)$	$D_1(L_m)$	$D_2(L_m)$	...	$D_k(L_m)$	Label-T_2
$pK(L_{m+1}@T_2)$	$D_1(L_{m+1})$	$D_2(L_{m+1})$	...	$D_k(L_{m+1})$	Label-T_2
$pK(L_q@T_t)$	$D_1(L_q)$	$D_2(L_q)$	...	$D_k(L_q)$	Label-T_t

Fig. 2 Schemes of IT-enhanced approaches, including target-specific QSARs enhanced by injection of predicted affinities for other targets $t_i \neq T$, and IT-enhanced chemogenomics where target information is reduced to a target label, provided by the IDFP "barcode"



$pK(L_1@T_1)$	$D_1(L_1)$	$D_2(L_1)$	...	$D_k(L_1)$	$\Delta_1(T_1)$	$\Delta_2(T_1)$	$\Delta_p(T_1)$
$pK(L_2@T_1)$	$D_1(L_2)$	$D_2(L_2)$	...	$D_k(L_2)$	$\Delta_1(T_1)$	$\Delta_2(T_1)$	$\Delta_p(T_1)$
...	...	...	...	...	...	...	...
$pK(L_m@T_2)$	$D_1(L_m)$	$D_2(L_m)$	...	$D_k(L_m)$	$\Delta_1(T_2)$	$\Delta_2(T_2)$	$\Delta_p(T_2)$
$pK(L_{m+1}@T_2)$	$D_1(L_{m+1})$	$D_2(L_{m+1})$	...	$D_k(L_{m+1})$	$\Delta_1(T_2)$	$\Delta_2(T_2)$	$\Delta_p(T_2)$
$pK(L_q@T_t)$	$D_1(L_q)$	$D_2(L_q)$	...	$D_k(L_q)$	$\Delta_1(T_t)$	$\Delta_2(T_t)$	$\Delta_p(T_t)$

Fig. 3 Principle of EL-enabled models: targets are characterized by protein descriptor vectors Δ. This paper challeneges the degree to which descriptors are carriers of important information

$\Theta^2 \geq 0.35$. This $\hat{B}QSAR$ matrix was, on one hand, concatenated to treePH03 descriptors, and, on the other, used as such, as an alternative stand-alone DS in GA selection.

5. `WE-IT`—Weak explicit inductive transfer differs from SE-IT by the use of unbiased predictors as additional descriptors (Fig. 2b). Refer to "Consensus and unbiased predictors" section for the difference between biased and unbiased estimators.
6. `ITCG`—The IT-enhanced CG approach (Fig. 2c) is a CG model trained on the entire multi-endpoint activity profile, providing only target identities (**IDFP**).
7. `ELP`, `ELSeq` and `ELSim`—EL-enabled CG experiments, using respective genuine protein descriptors Δ = {PROFEAT, sequence counts and SIMFP} (Fig. 3).

Model equivalence indices

The best models of each category listed in Table 2 will be compared to each other, in order to outline what, if anything, is specific to EL-enabled models with respect to IT-enhanced approaches and basic QSAR. Comparison will be consistently carried out in terms of the unbiased predictors $p\tilde{K}_i(M,T)$. Let the absolute prediction error, or `residual`, of a current ligand–target pair M,T committed by the best model representing an approach μ from Table 2 be denoted by $\epsilon^{\mu}(M,T)$

$$\epsilon^{\mu}(M,T) = \left|pK_i(M,T) - p\tilde{K}_i(M,T)\right| \qquad (5)$$

It is straightforward to calculate the root-mean-square of errors from Eq. (5) over all ligands of a target T, in order to obtain $RMSE^{\mu}_{T}$, respectively over the entire set of M,T pairs, leading to $RMSE^{\mu}$. The latter represents the cross-validated predictive power of the `strategy` μ over the entire set of 9,462 experimental entries. Strategies with significantly lower $RMSE^{\mu}$ are better. However, two strategies at roughly equal $RMSE$ are not necessarily equivalent. Equivalence, i.e. redundancy, of two models means that they commit comparable, or at least correlated errors for all predicted items. It is typically [17] assessed by estimating the dot product of residuals.

Inter-residual correlation scores ρ of two strategies (μ, v) are calculated after recentering these on the baseline BQSAR values, as in Eq. (6).

$$\rho(\mu, v) = \frac{\sum_{(M,T)}[\epsilon^{\mu}(M,T) - \epsilon^{BQSAR}(M,T)] \times [\epsilon^{v}(M,T) - \epsilon^{BQSAR}(M,T)]}{\sqrt{\sum_{(M,T)}[\epsilon^{\mu}(M,T) - \epsilon^{BQSAR}(M,T)]^2 \times \sum_{(M,T)}[\epsilon^{v}(M,T) - \epsilon^{BQSAR}(M,T)]^2}} \tag{6}$$

$$\pi(\mu,v) = \frac{\sum_T\left[RMSE_T^{\mu} - RMSE_T^{BQSAR}\right] \times \left[RMSE_T^{v} - RMSE_T^{BQSAR}\right]}{\sqrt{\sum_T\left[RMSE_T^{\mu} - RMSE_T^{BQSAR}\right]^2 \times \sum_T\left[RMSE_T^{v} - RMSE_T^{BQSAR}\right]^2}} \tag{7}$$

In this work, two strategies are redundant if they have similar impacts on the prediction errors of each item, compared to the basis strategy; i.e. if their BQSAR-centered vectors of residuals are collinear.

Alternatively, it is also advisable to monitor the differential impact of each strategy on the average prediction accuracy for each target, as the cosine of the angle formed but the N_T dimensional vectors of RMSE value *shifts*—see Eq. (7). Let this alternative target-based correlation score be denoted $\pi(\mu, v)$.

ρ and π thus represent metrics in a strategy space, in which the approaches from Table 2 can be located.

Deorphanization challenges

Twelve of the initially considered 31 GPCRs were removed from the initial pharmacological profile, and considered orphans. The remaining 19 proteins and their associated ligand sets (those which also bind the 12 orphans not excluded) were used to retrain models. The choice of the presumed orphans was done such as to encompass some examples having quite near analogs within the 19 training targets, while others were removed together with all their close analogs. Table 3 displays the fake orphans, together with the percentage of own ligands still in the training set, and the SIMPF versus top analogs amongst training proteins.

An entire subfamily (Muscarninc receptors) was thus included among the orphans. For some, most of their ligands are still present at the training stage, in association with one or several of the kept targets. Others have compound sets not overlapping with training ligands.

Typically, successful deorphanization amounts to the discovery of some actives for a so-far ligandless target. Therefore, active/inactive classifier models would suffice. Here, the challenge is not only to pinpoint potential binders, but to actually predict their pK_i values. For practical purposes, a method correctly ranking would-be binders by predicted potency would count as a success—therefore, herein we will use the Pearson correlation coefficient $R^2(p\hat{K}_i, pK_i)$ between predicted and experimental values for the orphan target as success score, rather than the stricter RMSE.

Theoretically, single endpoint QSARs are not extrapolable to other targets. Practically, however, the unknown model of the orphan T may be simply replaced by a model of a related t within the training set: $p\hat{K}_i(m,T) \equiv p\hat{K}_i(m,t)$. In this benchmarking study, the substitute model is the best model minimizing the $RMSE(p\hat{K}_i(m,t), pK_i(m,T))|_m$. This upper limit of the success of such `deorphanization by substitution` strategy serves as an estimator of default ‘deorphanizability’ in the given context.

SE-IT approaches were also employed as substitutes. Unlike BQSAR models, these had to be fully refitted, after removal of the predictors of pK_i values for orphan targets from the enhanced activity-descriptor matrices. FQ, QSAR and WE-IT strategies were not part of the deorphanization challenge.

The training set of CG models was reconstructed in ignoring entries from orphan targets. Prior to this, IDFP and SIMFP were edited: their size was cut to $N_t = 19$, as all cross-terms referencing orphan targets were deleted. ELP and ELSeq required no editing of the protein descriptors. Respective models were rebuilt on the basis of optimal libsvm setups obtained, for each strategy, at the previous learning stage over the complete profile.

Results and discussion

This section begins with a discussion of the herein advocated SVR-based CG model building scheme, in the context of other widely used CG approaches. This is followed by a brief debate on the rationale for and expectations from each of the learning strategies given in Table 2. Eventually, the predictive power of the approaches and their similarity in strategy space is discussed, first in the cross-validated model building, then in the deorphanization context.

Table 2 Various computational strategies used in this work

Strategy	Training sets	Ligand descriptors	Target descriptors	Remarks
FQ	Ligands of all targets	All[a]	–	Family QSAR: 'GPCR-likeness', ignoring targets
BQSAR	Ligands of target T	All[a]	–	Default QSAR model of each target
QSAR	Ligands of target T	treePH03	–	Default QSAR model in the treePH03 space
SE-IT	Ligands of target T	treePH03 + $\hat{B}QSAR_t\|_{t\neq T}$	–	Strong explicit inductive transfer
WE-IT	Ligands of target T	treePH03 + $\tilde{B}QSAR_t\|_{t\neq T}$	–	Weak explicit inductive transfer
ITCG	Ligand–target pairs	All[a]	IDFP	IT-enhanced CG, using IDFP for target labeling
ELSim	Ligand–target pairs	All[a]	SIMFP	EL-enabled model based on target similarity fingerprints
ELP	Ligand–target pairs	All[a,b]	PROFEAT	EL-enabled model using PROFEAT target descriptors
ELSeq	Ligand–target pairs	treePH03	seqAA	EL-enabled model using aminoacid sequence counts

[a] All the considered ligand descriptors (FPT1 seqSY37 seqPH37 aabPH02 treeSY03 treePH03) were allowed to compete (as such, or concatenated to target descriptors) in the Darwinian model optimization procedure

[b] A second run, based only on treePH03 descriptors for ligands—the winners of the all-descriptor evolutionary challenge—was rerun in order to ensure that no better model could be found when shrinking the search space in focusing on optimal descriptors

Support vector regression with concatenated descriptors

Concatenation of ligand and protein descriptors is certainly the most simple and straightforward approach [3, 4, 50]. Let us denote a ligand–target complex as $m:t$. Under the herein used concatenation hypothesis, $m:t$ will be described by a vector $\mathbf{m}:\mathbf{t}|_{concat}$ as defined below:

$$\mathbf{m}:\mathbf{t}|_{concat} = D_1(m), D_2(m), \ldots, \Delta_1(t), \Delta_2(t), \ldots \tag{8}$$

Authors [23, 25, 50] felt that cross terms $D_i(M)\Delta_j(T)$ should be explicitly added to the descriptor vector of complexes, allegedly representing ligand–protein feature interactions.

Table 3 List of presumed orphan targets, reporting the degree of ligand set overlap with training data, and the the closest analogs still in training set (relatedness—in terms of SIMFP—in parentheses)

Presumed orphan	% Ligs. in training set	Related targets in training set (SIMFP)
Alpha-1b adrenergic	99.29	Alpha-1d adrenergic (0.37)
		Alpha-1a adrenergic (0.22)
		Dopamine D2 (0.20)
Alpha-2b adrenergic	98.94	Alpha-2a adrenergic (0.58)
		Alpha-1d adrenergic (0.12)
		Alpha-1a adrenergic (0.06)
Alpha-2c adrenergic	90.76	Alpha-1d adrenergic (0.28)
		Serotonin 2c 5-HT2c (0.17)
		Serotonin 2b 5-HT2b (0.14)
Dopamine D5	98.98	Dopamine D1 (0.74)
		Serotonin 2a 5-HT2a (0.57)
		Dopamine D2 (0.28)
Melatonin 1A	100.00	Melatonin 1B (0.30)
Muscarinic acetylcholine M1	6.62	Serotonin 2a 5-HT2a (0.07)
Muscarinic acetylcholine M2	1.60	–
Muscarinic acetylcholine M3	4.70	Beta-2 adrenergic (0.47)
Muscarinic acetylcholine M4	7.46	–
Muscarinic acetylcholine M5	7.84	–
Serotonin 1b (5-HT1b)	90.58	Serotonin 1d 5-HT1d (0.33)
		Serotonin 1a 5-HT1a (0.27)
		Serotonin 6 5-HT6 (0.22)
Serotonin 5a (5-HT5a)	54.43	Serotonin 7 5-HT7 (0.28)
		Serotonin 6 5-HT6 (0.15)
		Serotonin 2c 5-HT2c (0.05)

As mentioned in "Introduction" section, the "kernel trick" [21, 22] was specifically introduced to support model building in descriptor spaces of arbitrarily high dimensionality, accounting for all cross terms in CG models. This accounting is implicit, i.e. cross-terms need not to be individually addressed because of the simple decomposition of the similarity score between two complexes (kernel) as a product of ligand–ligand and target–target similarities, as already highlighted in Eq. (3).

It is straightforward to show that, under the herein employed concatenation hypothesis, radial basis function (RBF)-based kernels also satisfy relation (3).

Therefore, it appears that the choice of describing complexes by simple concatenation of partner descriptors can be equivalent to the kernel product approach, assuming the appropriate choice of the kernel form (RBF, for instance). The observations reported here are therefore likely to generally apply to kernel- or descriptor-based non-linear CG experiments.

It was observed that the RBF kernel systematically emerged as winner in the evolutionary process of selecting optimal model building configurations for the libsvm regression tool, for all the CG models (for some of the individual QSAR models, the linear kernel was preferred). We interpret this as a consequence of the above-mentioned ability of RBF kernel to implicitly emulate contributions of protein–ligand descriptor cross terms.

Furthermore, for all but one case (ELSeq), Darwinian evolution coherently rediscovered that the min/max scaled treePH03 ligand descriptors concatenated to the current protein fingerprint, is the best model provider. This was expected, since in non-rescaled concatenated vectors, ligand and protein terms may span different orders of magnitude, and thus cause artifacts.

Rationale of the modeling strategies

This subsection briefly revisits each model building strategy, in order to justify its inclusion in this benchmarking study and, if appropriate, position it in the context of similar work reported in literature.

Basic QSAR provides a must-have baseline to compare CG models against. It serves as a generic measure of modelability of data. Individual QSAR models may be each built in a different, best-suited DS. Simultaneous multi-endpoint learning, irrespectively of employed kernels, must happen in a common ligand-protein DS. Therefore, an explicit distinction between optimal BQSAR, in the best suited DS, and QSAR models confined to the consensus DS (see Table 2) are explicitly benchmarked here.

Classical, ligand-based QSAR approaches cannot formally serve to predict affinities for targets other than they were calibrated for. However, single-endpoint QSAR models may be successful in finding ligands of an orphan target, if this is strongly related to the training target. Basic "substitution" of the predictor for the orphan by the models of the training targets might reveal that at least one of these would, eventually, correctly prioritize binders to the orphan at the top of the list of candidates sorted by decreasing predicted activity on the training target. Deorphanization by substitution is thus a measure of the intrinsic degree of difficulty of such a challenge. It should become, in our opinion, a compulsory test accompanying such studies.

Another important benchmarking baseline is provided by the FQ strategy. As target information is withheld, a same set of descriptors $\mathbf{D}(M)$ may occur several times, associated to pK_i values stemming from different targets. However, such a model, aimed at learning the common "family-specific" signature of GPCR ligands, may nevertheless reach high statistical robustness levels. If there is little or no overlap of ligand subsets associated to each target, or, alternatively, if most compounds happen to have roughly equal activity levels on all the targets they were tested against, then a multi-domain model building method may reach a performance level close to basic QSAR. Assessing this generic model is therefore also a must: if easy to cross-validate, it signals that training data are biased and do not span sufficient chemical and target diversity in order to support CG analysis.

A further issue being dealt with here is the question whether the nature of IT-enhancement is dependent on the single- or multi-endpoint-nature of the learning strategy. IT may be exploited in both contexts: hence the need for explicit comparison between SE/WE-IT and ITCG. ITCG, reading the entire activity profile, has simultaneous access to all the experimental $pK_i(m,t)$. In SE-IT and WE-IT models of a target *T*, predicted affinities of its ligands with respect to all the other endpoints $t \neq T$ will be "given a chance" to illustrate themselves as useful players in the *T* QSAR model. In SE-IT, ligands *m* of *T* that also happened to be members of the *t* training set will be represented by quite accurate activity estimates $p\hat{K}_i(m,t)$ close to fitted values. By contrast, in the weak hypothesis, the $p\tilde{K}_i(m,t)$ are more noisy.

ITCG, the multiple-endpoint IT-enhanced learning strategy, corresponds in many respects to previously reported models [21, 22]. Since herein reported ITCG models rely on the RBF kernel of libsvm, kernel product decomposition (3) applies for the concatenated ligand–protein vector. In this context, $\mathcal{K}^{prot}(t,T) = exp(-4\gamma)$ if $t \neq T$, else $\mathcal{K}^{prot}(t,t) = 1.0$. This is nothing but the two-state "multitask" kernel used in the cited works, equal to some empirical value below 1.0 (typically 0.5) if $t \neq T$, and 1.0 otherwise.

Two out of three EL-enabled approaches, ELP and ELSeq, rely on classical sequence-based protein fingerprints. These are global descriptors of protein sequences, and are therefore at risk to "drown" the key information related to the actual binding sites amongst contributions stemming from irrelevant subdomains. However, descriptors such as PROFEAT or seqAA are important because they can be generated for any protein of known sequence, including orphan targets with undefined binding sites.

By contrast, SIMFP have been designed to target the specific ligand-binding behavior of proteins. These are experimentally measured, rather than computed, descriptors of affinity covariance with respect to common ligands.

Table 4 RMSE values of each strategy with respect to the compound sets of every target, and global RMSE over the entire set

Target	ELSeq	ELP	ELSim	ITCG	FQ	SE-IT	WE-IT	QSAR	BQSAR
Alpha-1a adrenergic	0.84	0.82	0.82	0.87	1.29	0.85	0.84	0.85	0.85
Alpha-1b adrenergic	0.47	0.51	0.47	0.50	0.86	0.52	0.59	0.59	0.60
Alpha-1d adrenergic	0.57	0.64	0.54	0.60	0.87	0.64	0.74	0.72	0.72
Alpha-2a adrenergic	0.61	0.58	0.55	0.65	0.86	0.55	0.58	0.60	0.60
Alpha-2b adrenergic	0.45	0.47	0.42	0.54	0.72	0.33	0.41	0.53	0.45
Alpha-2c adrenergic	0.67	0.62	0.66	0.66	1.26	0.61	0.66	0.66	0.66
Beta-1 adrenergic	0.59	0.66	0.58	0.57	0.85	0.59	0.70	0.74	0.66
Beta-2 adrenergic	0.71	0.75	0.69	0.71	0.92	0.69	0.75	0.78	0.72
Beta-3 adrenergic	0.62	0.56	0.65	0.68	0.95	0.54	0.61	0.58	0.53
Dopamine D1	0.60	0.54	0.57	0.59	0.83	0.49	0.56	0.56	0.55
Dopamine D2	0.58	0.58	0.57	0.58	0.84	0.58	0.60	0.59	0.60
Dopamine D3	0.62	0.66	0.61	0.63	1.02	0.69	0.69	0.69	0.69
Dopamine D4	0.60	0.65	0.63	0.68	1.28	0.64	0.63	0.63	0.64
Dopamine D5	0.64	0.67	0.56	0.65	0.88	0.61	0.78	0.78	0.73
Melatonin 1A	0.74	0.68	0.70	0.65	0.93	0.64	0.70	0.71	0.70
Melatonin 1B	0.71	0.79	0.75	0.74	1.05	0.82	0.79	0.79	0.80
Muscarinic M1	0.77	0.95	0.80	0.83	1.04	0.84	0.91	0.93	0.93
Muscarinic M2	0.54	0.67	0.56	0.56	0.81	0.65	0.83	0.86	0.81
Muscarinic M3	0.65	0.74	0.69	0.71	1.03	0.71	0.80	0.82	0.78
Muscarinic M4	0.39	0.53	0.37	0.39	0.58	0.47	0.72	0.71	0.70
Muscarinic M5	0.39	0.47	0.42	0.49	0.95	0.61	0.86	0.82	0.72
Serotonin 1a (5-HT1a)	0.64	0.61	0.63	0.63	0.73	0.61	0.63	0.62	0.61
Serotonin 1b (5-HT1b)	0.63	0.71	0.62	0.66	0.95	0.69	0.80	0.81	0.78
Serotonin 1d (5-HT1d)	0.67	0.77	0.73	0.76	0.99	0.67	0.84	0.83	0.75
Serotonin 2a (5-HT2a)	0.68	0.66	0.68	0.70	1.00	0.68	0.69	0.69	0.67
Serotonin 2b (5-HT2b)	0.59	0.56	0.56	0.57	0.77	0.57	0.63	0.63	0.62
Serotonin 2c (5-HT2c)	0.62	0.65	0.63	0.66	0.94	0.64	0.68	0.67	0.67
Serotonin 4 (5-HT4)	0.67	0.67	0.64	0.69	1.13	0.64	0.65	0.64	0.55
Serotonin 5a (5-HT5a)	0.78	0.74	0.80	0.78	0.92	0.72	0.73	0.72	0.72
Serotonin 6 (5-HT6)	0.59	0.58	0.59	0.60	0.68	0.55	0.55	0.55	0.55
Serotonin 7 (5-HT7)	0.64	0.63	0.64	0.65	0.80	0.64	0.64	0.64	0.61
All targets	0.63	0.64	0.63	0.65	0.92	0.63	0.67	0.67	0.66

Their ab initio prediction based on protein sequences/structures would be a quite challenging task. Albeit not pin-pointing the actual ligand–target interaction points, and not explicitly referring to the involved active site residues, an affinity covariance score matrix like SIMFP, might, in principle, traced back to a set of mechanistically relevant protein descriptors, by embedding [31]. To our knowledge, none of the computationally available protein descriptors accurately encodes the information needed to predict affinity covariance of proteins. SIMFP were employed in this study in order to benchmark the "optimistic hypothesis" of CG simulations benefiting from a quasi-ideal scenario in terms of protein descriptor quality. In practice, their real information content is limited and biased by the actual choice of common ligands having served to derive them. In previous publications, their closest equivalent would be "hierarchy" protein kernels [21, 22, 34], conveying a knowledge-based, empirical measure of relatedness of two proteins. Nevertheless, ELSim is thus expected to be the best-suited candidate for triggering an actual EL-enhancement. Whilst applicable to simulated deorphanization challenges, ELSim is useless in genuine, prospective deorphanization: SIMFPs cannot be built for targets other than well-studied, ligand-rich proteins.

treePH03: the consensus chemical space

Strategies BQSAR, FQ, ITCG, ELSim and ELP, all enabled to freely pick an optimal ligand DS out of given options, were chronologically the first performed. FQ, ITCG, ELSim and ELP, as well as a significant number of target-specific BQSAR models independently favored the treePH03 DS. Subsequent runs were therefore confined to this consensus space, as common basis for inter-strategy comparison. BQSAR was picked as provider of baseline residual values in Eq. (6). It has the advantage of being rather centrally located in strategy space: neither the most, nor the least potent.

Overall performance in cross-validated model building

A glance at the last column of Table 4 shows no outstanding overall advantage of EL-enabled and IT-enhanced methods in terms of average error over the entire compound set. At overall standard deviation of the pK_i values of 1.19, $0.63 \leq$ RMSE ≤ 0.67 correspond to determination coefficients between 0.72 and 0.68 and compare favorably to SVR-based CG models in an analogous study within the GPCR family [3].

Controlled by ligand-rich targets, not benefiting from IT enhancements, global RMSE is fairly strategy-independent—with one notable exception, the FQ model. This latter clearly fails to cross-validate properly, thus proving that the chosen data set is not trivial to model. FQ will therefore be ignored in the following analysis. However, it is worthwhile mentioning that a second set of kinase binders, alternatively considered for this study, was eventually rejected because it failed to pass the FQ test, i.e. scored FQ results comparable to those of classical QSAR (results not shown). This basically meant that, within that set, selective ligands were either rare, or their cross-testing results on the other targets were not reported. All that could have been learned from that collection would have been the common features of the "average" ATP-like kinase binder.

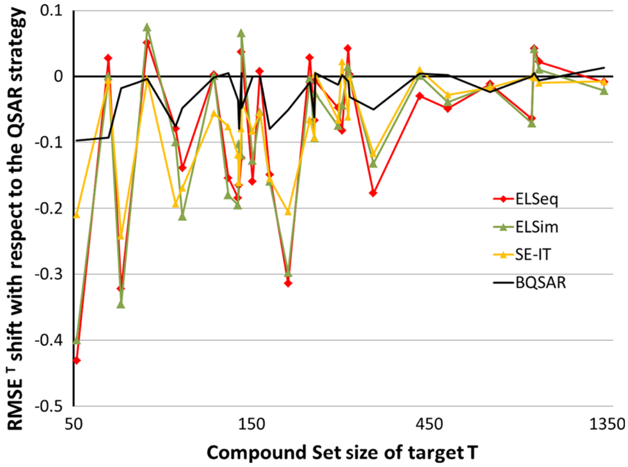


Fig. 4 Relative shift of the $RMSE_T$ values reported by the displayed strategies, with respect to the $RMSE$ of the consensus DS-based QSAR strategy, on Y, as a function of target set sizes—on X, logarithmic scale

Figure 4 displays, for four representative strategies μ (others left out in order to avoid overcrowding), the shift of the per-target RMSE values with respect to treePH03-based QSAR, $RMSE_T^{\mu} - RMSE_T^{QSAR}$, plotted against the ligand subset size of each t. Significant gains in model quality are more often seen with smaller sets. Not all the smaller sets report gains, however. ELSeq and ELSim are seen to perform quasi-identically, specifically enhancing predictions of the same targets by identical amounts. ELP (not shown) also draws a pattern neatly overlapping to these two. SE-IT and ITCG (not shown) are also significantly, but slightly less covariant with ELSeq and ELSim.

The curves of all the IT-enhanced and EL-enabled methods have a same shape. By contrast, the difference between BQSAR and QSAR displays the gains due to choosing the ligand DS to fit the specific requirements of each target. Occasional losses due to adoption of a unique DS are significantly less important than the gains stemming from IT. The most important IT enhancements are seen to occur for muscarinic receptors M4 and M5, with small associated sets of 51 and 67 reported ligands, and well surrounded in target space by a series of the other three, significantly more data-rich muscarinics.

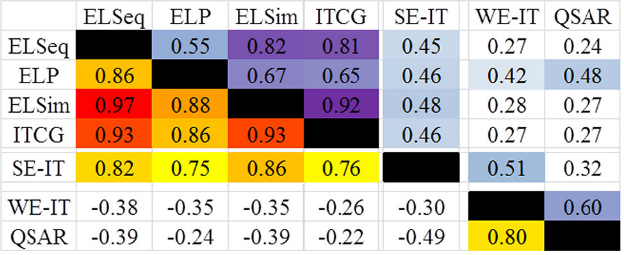
Actually, the comparison of Fig. 4 to the conceptually similar plot of relative AUC enhancements [22] reveals a much more robust effect, extending to set sizes of the order of 10^2. By contrast, the cited work witnessed significant enhancement only for minimalistic compound set sizes of ≤ 50 (≤ 20 for GPCRs).

Error pattern analysis in cross-validated model building: Are these strategies different?

Figure 5 displays, in a diagonally split matrix, the relative degrees of similarity of cross-validated prediction results returned by each strategy. The upper-right triangle reports ρ correlation degrees of residual error shifts with respect to BQSAR, the lower-left monitors π covariance of relative changes in RMSE values per target.

The figure clearly unveils, with respect to either π or ρ metrics, that the strategies can basically be grouped in two distinct, well-defined blocks. The residual pattern shift of QSAR versus BQSAR is the signature of the coerced move to a unique ligand DS, as required for multi-endpoint learning. While the BQSAR → QSAR move in strategy space does not significantly increase the global RMSE, in terms of individual prediction patterns it nevertheless produces a rather specific signature, largely uncorrelated to others—except WE-IT.

WE-IT is a 'diluted' version of explicit IT, corresponding to the worst-scenario expectation of learning



	ELSeq	ELP	ELSim	ITCG	SE-IT	WE-IT	QSAR
ELSeq		0.55	0.82	0.81	0.45	0.27	0.24
ELP	0.86		0.67	0.65	0.46	0.42	0.48
ELSim	0.97	0.88		0.92	0.48	0.28	0.27
ITCG	0.93	0.86	0.93		0.46	0.27	0.27
SE-IT	0.82	0.75	0.86	0.76		0.51	0.32
WE-IT	-0.38	-0.35	-0.35	-0.26	-0.30		0.60
QSAR	-0.39	-0.24	-0.39	-0.22	-0.49	0.80	

Fig. 5 Clustering of strategies in terms of the two considered model equivalence indices ρ and π. Cells (μ, ν) host *color-coded* values of indices $\rho(\mu, \nu)$ and $\pi(\mu, \nu)$ respectively. Since both ρ and π matrices are symmetrical with *unit diagonals*, they are being merged in this figure: the *upper-right* half reports $\rho(\mu, \nu)$, the correlation levels of BQSAR-centered residual errors over the entire data set, color-coded in *blue–violet* over the significant range between 0.4 and 1. The *lower-left triangle* reports $\pi(\mu, \nu)$ values, the correlation levels of BQSAR-centered RMSE shifts for each target, color-coded in *yellow–red*

from an activity profile with mutually disjoint ligand subsets for each target. Knowledge is transferred via unbiased, error-prone external predictions of pK_i values. The noise affecting the latter makes them, indeed, less useful descriptors in explicit IT models. Therefore, WE-IT, a single-point strategy in treePH03, like QSAR, does not notably diverge from this latter. According to the π metric, however, WE-IT is seen to represent a bridging element between QSAR and SE-IT.

ITCG and EL-enabled techniques have direct access to all the experimental values over all endpoints, while SE-IT injects rather accurate estimators of pK_i values of the shared ligands in data sets. Residual shift patterns clearly locate the single-endpoint SE-IT within the multi-endpoint method family, albeit at the cluster border to basic QSAR. However, this method is the provider of many amongst the most accurate per-target results herein reported: Alpha-2B (RMSE = 0.33; next best at 0.41), Dopamine D1 (RMSE = 0.49, next best at 0.54).

The clustering observed in Fig. 5 primarily reflects whether IT is enabled or not. The specific implementation of IT—explicit transfer in single-point models versus the CG approaches (all IT-enhanced, some EL-enabled)—has a minor impact on predictions. If anything, IT enhancement seems less marked in ELP—probably due to noise caused by some irrelevant PROFEAT terms.

There is clearly no specific signature difference between EL-enabled and IT-enhanced approaches. The approach less well fitting within the tight cluster of the four multi-endpoint approaches is ELP—unfortunately, standing out due to its slightly decreased performance. This notwithstanding, EL-Sim, the strategy based on the most promising protein descriptors and a priori though to be the most likely to reach actual EL-enhanced status, is often seen to win, by sometimes significant margins, in terms of per-target RMSE.

However, as already hinted, cross-validated model building are not sufficient proof in favor of EL-enhancement.

Outlier analysis

As a direct consequence of above-mentioned covariance of error patterns, the sets of ligand–target complexes that were mispredicted by the different methods tend to overlap significantly. Irrespective of the nature of the method (FQ excluded), the number of (cross-validation) $p\tilde{K}_i$ predictions affected by errors of two order of magnitude or more, all ligands of all targets confounded, typically lies between 50 and 90, e.g. roughly between 5 and 10 % of the entire set of data. Twenty ligand–target pairs emerge as consensus "outliers" (in the above-mentioned sense of prediction errors beyond 2.0) simultaneously mispredicted by all the methods. They thus typically represent more that 25 % of any method-specific outlier subsets. As expected, outlier subsets of specific approaches seen to be closely clustered in strategy space overlap significantly more, with up to 80 % of common instances.

The structures of these 20 consensus outliers are given in Table 5. They include both over- and under-predicted affinities. Note that, however, in a practical virtual screening scenario based on these models, underpredictions would in most cases not have been penalizing, since returned predictions were better than micromolar, i.e. above the typical selection threshold.

Outliers do not appear to be preferentially associated to any target, or to any specific structural signature of the ligand. At best, the last two amongst the four D3 inhibitors display some similarities: both are chemically similar, highly flexible and large compounds being underpredicted by the approaches. Intriguingly, only five of the involved ligands are chiral, thus potentially explainable by the failure to incorporate chirality-related information in the used molecular descriptors. The 20 entries have been manually cross-checked against the ChEMBL database, in order to exclude potential glitches in the original data importation, standardization and outlier analysis process. In light of the robustness of the modeling approaches, this consensual highlighting of a set of apparently unrelated compounds may hint towards actual problems with the reported affinity values [35].

Deorphanization challenge results

The fundamental distinction between EL- and IT-driven enhancement may only become apparent in deorphanization challenges, as argued in "Introduction" section. However, prior to this authors need to thoroughly whether the attempted deorphanization task actually represents a challenge.

Table 5 Structures and identifiers of the 20 outliers consensually mispredicted by more than two orders of magnitude by all of the models

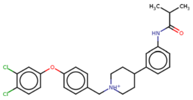
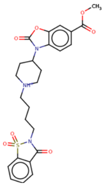
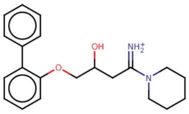
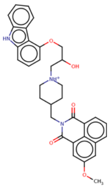
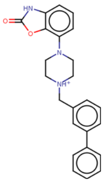
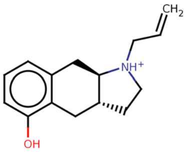
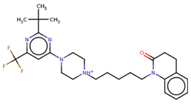
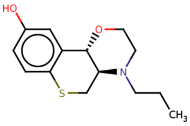
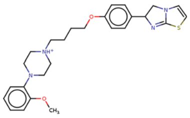
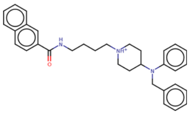
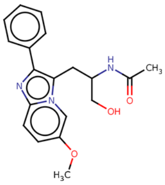
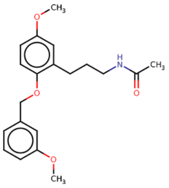
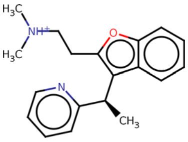
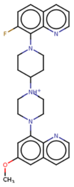
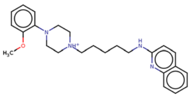
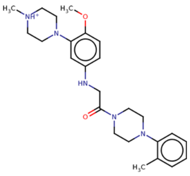
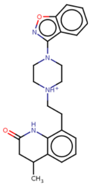
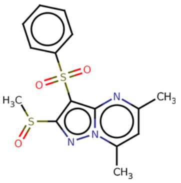
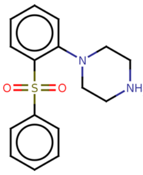
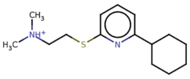
Structure	Expt. pK_i	Predicted	Compound CHEMBL ID	Target CHEMBL ID	Target name
	4.47	6.77 ± 0.11	389129	229	Alpha-1a adrenergic
	5.99	8.19 ± 0.15	311290	229	Alpha-1a adrenergic
	5.60	7.96 ± 0.13	181583	229	Alpha-1a adrenergic
	10.00	7.30 ± 0.19	1084707	213	Beta-1 adrenergic
	10.20	7.54 ± 0.12	218166	217	Dopamine D2
	9.52	6.86 ± 0.07	11363	217	Dopamine D2
	5.39	7.95 ± 0.11	201506	234	Dopamine D3
	5.52	8.50 ± 0.09	98197	234	Dopamine D3
	9.92	7.54 ± 0.20	464395	234	Dopamine D3
	9.52	7.39 ± 0.10	239923	234	Dopamine D3
	5.10	7.62 ± 0.20	1823018	1946	Melatonin 1B

Table 5 continued

Structure	Expt. pK_i	Predicted	Compound CHEMBL ID	Target CHEMBL ID	Target name
	12.16	8.53 ± 0.27	1092646	1946	Melatonin 1B
	9.49	6.30 ± 0.40	563920	216	Muscarinic M1
	6.56	8.86 ± 0.10	1096530	214	Serotonin 1a
	10.52	7.93 ± 0.19	271987	214	Serotonin 1a
	4.38	7.08 ± 0.41	110601	1983	Serotonin 1d
	6.72	9.58 ± 0.08	382748	224	Serotonin 2a
	5.53	7.75 ± 0.07	1762577	3371	Serotonin 6
	5.40	7.92 ± 0.32	209893	3371	Serotonin 6
	9.15	6.84 ± 0.20	158402	3155	Serotonin 7

Predicted values report mean and standard deviations over predictions from all the strategies except FQ

Deorphanization by substitution

The fact that EL may in principle extrapolate the model weighing factors for an orphan target, while pure IT cannot refit them, becomes inconsequential if there is no need to refit. A model trained on a very close analogue of the orphan may "happen" to return usefully accurate estimates of the affinity for the orphan. Such "deorphanization by substitution" provides a default measure of how easy candidate targets can be deorphanized. Here, substitutions

Table 6 Substitute targets from the training set providing single-endpoint BQSAR and SE-IT models, respectively, that best mimic the orphan's structure-activity model

Presumed orphan T	BQSAR substitute t	BQSAR R^2	SE-IT Substitute t'	SE-IT R^2
Alpha-1b adrenergic	Beta-2 (0.00/1)	0.10	Dopamine D3 (0.00/0)	0.06
Alpha-2b adrenergic	Alpha-2a (0.58/60)	0.58	Alpha-2a (0.58/60)	0.58
Alpha-2c adrenergic	Serotonin 1d (0.00/5)	0.04	Serotonin 1d (0.00/5)	0.09
Dopamine D5	Dopamine D1 (0.74/25)	0.67	Dopamine D1 (0.74/25)	0.67
Melatonin 1A	Melatonin 1B (0.30/82)	0.28	Melatonin 1B (0.30/82)	0.30
Muscarinic M1	Beta-2 (0.00/0)	0.05	Serotonin 6 (0.00/0)	0.01
Muscarinic M2	Serotonin 2a (0.00/0)	0.08	Melatonin 1B (0.00/0)	0.05
Muscarinic M3	Serotonin 2a (0.00/0)	0.12	Serotonin 2a (0.00/0)	0.12
Muscarinic M4	Serotonin 1d (0.00/1)	0.33	Serotonin 6 (0.00/0)	0.09
Muscarinic M5	Beta-3 (0.00/0)	0.04	Serotonin 2c (0.00/0)	0.21
Serotonin 1b	Serotonin 1d (0.33/66)	0.16	Alpha-1d (0.00/1)	0.25
Serotonin 5a	Serotonin 7 (0.28/15)	0.19	Alpha-2a (0.00/2)	0.18

In parentheses (SIMFP relatedness to the presumed orphan T/percentage of T ligands also part of the t set, therefore contributing to fit the substitute model $p\hat{K}_i(m,t)$). The correlation between $p\hat{K}_i(m,t)$ and experimental $pK_i(m,T)$ is reported in terms of Pearson R^2 values

by the BQSAR, respectively SE-IT models best predicting the pK_i values of the orphan (in terms of Pearson correlation scores) are reported in Table 6.

Recall that SIMFP represent the observed likelihood that a pK_i value measured with respect to t is a meaningful estimator of the affinity for the so-far assumed ligandless T. If the presumed orphan's ligand set significantly overlaps with the one of t—or, at least, with the applicability domain of $p\hat{K}_i(m,t)$—then $p\hat{K}_i(m,t)$ are accurate estimates pf $pK_i(m,t)$ and reported R^2 values tend towards SIMFP score levels. If reported R^2 values do not dramatically decrease with respect to SIMFP scores as ligand set overlap values diminish, this is indirect proof of excellent robustness and external prediction propensity of single endpoint models. Such is the case with the Alpha-2b versus Alpha-2a (no loss of correlation, at 0.58, in spite of the 40 % of external compounds) Dopamine D5 versus D1 (marginal loss from 0.74 to 0.67, at 75 % of external compounds) or Serotonin 5a versus 7 (a marginal 0.19 subsists out of the already low 0.28, at 85 % of external compounds).

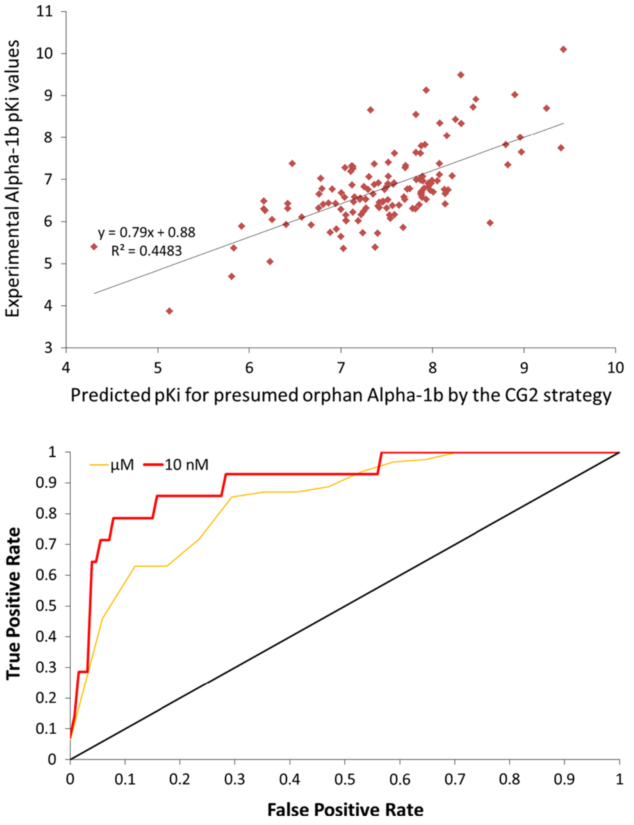


Fig. 6 *Above* experimental versus predicted pK_i plot for the ELSeq strategy with respect to presumed orphan Alpha-1b. *Below* corresponding ROC curve counting submicromolar (*orange curve*, AUC = 0.84), respectively better than 10 nM (*red*, AUC = 0.91) as 'active'

In only five cases (Alpha-1b, Alpha-2c, Muscarinic M3, Serotonin 1b and 5a—the latter two exceptions being specific to SE-IT models) the substitute model did not stem, as expected, from the target closest in SIMFP space—compare Tables 3 and 6. The first two exceptions are both related to the low quality of the Alpha-1d model, which is the common closest training set neighbor of both Alpha-1b and 2c. The third example occurring both with both BQSAR and SE-IT (M3) corresponds to quasi-null compound set overlap levels, likely representing an external prediction failure.

Occasionally—and this includes the two SE-IT-specific exceptions mentioned above—R^2 is significant (low, but within the range of typical SIMFP values) even though actual SIMFP failed to be estimated (assumed zero) due to insufficient ligand set overlap. For example, predicted Serotonin 1d affinities correlate with experimental muscarinic M4 affinities.

Only in two examples out of 12, deorphanization by substitution supports a quantitative prediction of affinities for the presumed orphan: Alpha-2b are well approximated

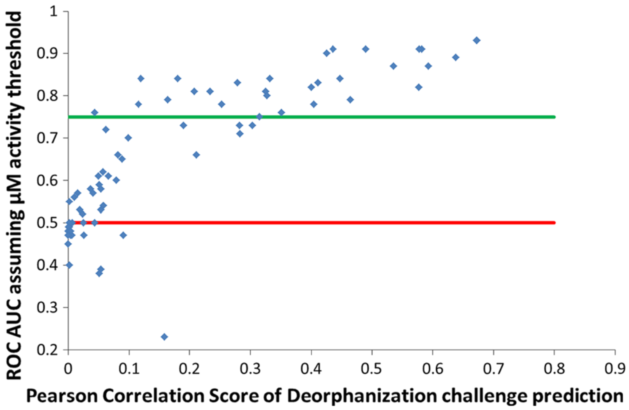


Fig. 7 Corresponding ROC AUC values taken at µM activity threshold against Pearson correlation coefficients given in Table 7, for each presumed orphan—strategy combination. The *red baseline* at 0.5 marks the random level, whereas the *green line* at 0.75 delimits the very robust results

Table 7 Pearson R^2 values of deorphanization challenge predictions, according to each considered strategy

Presumed orphan T	ELSeq	ELP	ELSim	ITCG	SE-IT	BQSAR
Alpha-1b (3: 0.37)	**0.448**	**0.325**	**0.411**	**0.400**	0.062	0.099
Alpha-2b (1: 0.58)	*0.425*	0.536	0.490	*0.436*	0.582	0.578
Alpha-2c (1: 0.28)	0.004	0.054	0.006	0.007	0.091	0.044
Dopamine D5 (3: 0.74)	0.577	0.593	0.638	*0.465*	0.673	0.672
Melatonin 1A (1: 0.30)	0.283	**0.404**	0.315	0.351	0.303	0.282
Muscarinic M1 (0: 0.00)	0.000	0.002	0.004	0.002	0.010	0.051
Muscarinic M2 (0: 0.00)	0.054	0.019	0.025	0.024	0.054	0.082
Muscarinic M3 (1: 0.47)	0.041	0.000	0.026	0.050	0.116	0.120
Muscarinic M4 (0: 0.00)	*0.001*	*0.159*	*0.000*	*0.037*	*0.089*	0.327
Muscarinic M5 (0: 0.00)	0.051	0.057	0.002	0.016	**0.211**	0.044
Serotonin 1b (3: 0.33)	0.234	0.208	**0.332**	0.279	0.253	0.164
Serotonin 5a (1: 0.28)	*0.058*	*0.002*	*0.079*	*0.066*	0.181	0.190

Bold/italicized entries—R^2 values above/below the BQSAR baseline score, by more than 0.1. Each presumed orphan T is characterized, in parentheses, by (the number of training set targets t with SIMFP similarity levels above 0.2: SIMFP value of closest analog)

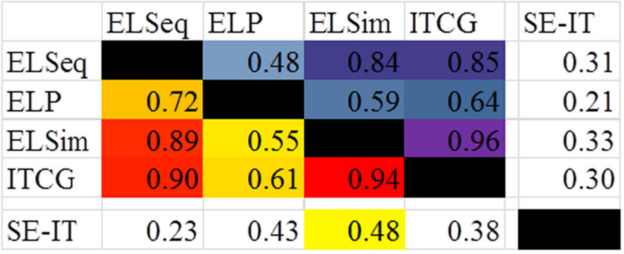
($R^2 = 0.6$) by Alpha-2a affinities, whilst D1 affinity is a robust predictor ($R^2 = 0.7$) of D5 binding. In the other cases, in as far as calculated R^2 values do not decrease well below SIMFP levels, direct target-to-target affinity extrapolation failure is likely an expression of actual target selectivity, not a consequence of individual model prediction inaccuracy. Also, SE-IT models have no significant advantage over BQSAR.

Deorphanization benchmark

As can be seen from Table 7, only one orphan target is significantly better predicted by the multi-endpoint methods, with respect to the already discussed substitution baseline values in the two right-most columns. Deorphanization of Alpha-1b by the four multi-endpoint methods is significantly more successful that the substitution-driven baseline attempt. Even though (Fig. 6) ELSeq-predicted pK_i values are not very accurate pK_i levels, they nevertheless allow for robust discrimination between potent and weak actives on this target.

On the absolute, Table 7 looks fairly discouraging, in stark contrast to robust ROC AUC values of 0.7–0.9 reported for deorphanization experiments [21, 22]. Current results are nevertheless not obviously worse. First, preceding works first tended to use a much more lenient leave-one-out deorphanization protocol, as already mentioned. Simultaneous removal of all the muscarinics from the training set was a risky bet leading to a not unexpected failure to deorphanize them. Furthermore, although the Pearson R^2 reported here is already a concession with respect to strict quantitative predicted-minus-experimental RMSE values, ROC AUC is much more lenient still. For comparative purposes, ROC curves were generated for all the quantitative predictions characterized by the R^2 values in Table 7, adopting micromolar activity as active/inactive threshold. As can be seen from Fig. 7, many low-R^2 experiments would go as robust successes by the AUC criterion, including, for example, the muscarinic target M4 (deorphanizd by substitution by Serotonin 1d, at ROC AUC = 0.80). However, the current paper focuses on quantitative prediction models, and will therefore keep its main focus on R^2 values.

Alpha-1b prediction is being enhanced by all multi-endpoint methods, including ITCG. Therefore, this is most likely due to a consensus effect, for Alpha-1b has three significantly related—and rather equally relevant—targets in the training set: Alpha-1d (SIMFP = 0.37), Alpha-1a and Dopamine D2 (both at SIMFP = 0.2). While neither of those individual models is a valid substitute for Alpha-1b activity, their consensus, in both plain ITCG and in two EL-enabled approaches, represent better approximations. As the 141 Alpha-1b ligands are largely shared by Alpha-1d (131), Alpha-1a (117) and Dopamine D2 (63), ITCG



	ELSeq	ELP	ELSim	ITCG	SE-IT
ELSeq		0.48	0.84	0.85	0.31
ELP	0.72		0.59	0.64	0.21
ELSim	0.89	0.55		0.96	0.33
ITCG	0.90	0.61	0.94		0.30
SE-IT	0.23	0.43	0.48	0.38	

Fig. 8 Clustering of strategies in the deorphanization challenge, based on ρ and π indices, now specifically calculated over orphan targets and their ligands only (see caption of Fig. 5 for details)

and ELSim predictions are, predominantly, extrapolations based on these values. Therefore, this would rank as a Class-III [7] problem (new protein, known training set ligands).

In ITCG, orphan *T*, described by a null fingerprint, is not properly defined and does not impact on calculated kernel values. Thus, ITCG indiscriminately includes, in the returned consensus, any $m - t$ ligand–target pair containing a ligand similar to the predicted one, irrespective of endpoint *t*. ITCG succeeded only because adrenergic ligands in the training set were predominantly associated to adrenergic and related dopaminergic receptors. Only 22 of the 141 Alpha-1b ligands are also associated to other training targets. Would all Alpha-1b ligands have been tested on other receptors of the training set, those likely irrelevant pK_i values would have mechanically impacted ITCG consensus predictions, with worsening results. Learning form an ideal, complete ligand × targets experimental profile would paradoxically penalize ITCG. Similarly, if a presumed orphan has only one (or several, yet not equally relevant) neighbors, the same consensus mechanism having contributed to the above-discussed deorphanization success may now lead to less good results than plain substitution. Indiscriminate dilution of a potent substitution model by less relevant data points causes a sensible quality loss of ITCG—see Alpha-2b and Dopamine D5.

By contrast, knowledge-enhanced ELSim systematically fares better or equally well than ITCG, by contrast to the other two EL-enabled approaches—expectedly so, in light of the serious drawbacks of full-sequence protein descriptors. In above-mentioned Alpha-2b and Dopamine D5 cases, ELSim manages to play down "contamination" effects significantly impacting ITCG. This observation might, at first sight, be thought of as evidence in favor of genuine EL-enhancement. Unfortunately, ELSim is *not* systematically outperforming the "null' substitution model hypothesis. Enhancement over ITCG—also see [21, 22] Vert's hierarchy kernel outperforming the "multitask approach"—is, unfortunately, not sufficient evidence in favor of EL-enhancement. Unambiguous proof in favor of EL enhancement also implies that the approach has succeeded in non-trivial deorphanization attempts, and clearly outperformed basic substitution.

Either way, no genuine EL enhancement beyond IT could be evidenced in this work. Successful deorphanizaion attempts were either trivial (i.e. achievable by using a single endpoint model of a very close receptor analogue included in the training set) or due to favorable consensus effects, and/or fortuitous incompleteness of the training set activity matrix. The π/ρ strategy space clusters (Fig. 8) for the deorphanization challenge again show a compact block regrouping both CG and MTL.

Conclusions

Chemogenomics should be considered as a generalization of QSAR, replacing the battery of single endpoint models by one multi-endpoint approach. Both prior chemogenomics modeling reports and this report have shown the improvements gained compared to classical QSAR. This does not reduce the strengths of QSAR, and researchers interested in a single target with a fixed ligand scaffold will continue to benefit from QSAR modeling.

In this work, we have addressed a key CG modeling question—whether or not the performance of the CG models is improved as a result of true protein descriptors, or if mere protein labels can provide equivalent modeling performance. Of course, no single study focusing on any given receptor family and choice of protein descriptors may provide a definite answer. The main goal of this work was to explicitly formulate this question, which had already been implicitly but, in our opinion, incompletely addressed in previous studies. Here, we formally distinguished between IT and explicit learning (EL) as possible competing scenarios contributing to the success of CG over classical QSAR. We have showed that simple comparison of cross-validated performance of two methods differing with respect to the information content of protein descriptors is not compelling evidence in favor of EL occurring when relevant protein descriptors are used. The extensive benchmarking protocol outlined here, including strategy clustering in prediction error spaces and systematic comparison of deorphanization results to the baseline receptor "substitution" strategy, revealed that sole improvement in cross-validation statistics is not a proof of EL. We would therefore like to encourage authors of further studies to fully apply the herein outlined protocol when searching for EL-enabled CG approaches.

Regression modeling was used to extensively investigate multiple variations of IT-enhanced and EL-enabled methods for their ability to predict quantitative pK_i values. The various approaches were assessed for their performance in a strategy space, which allowed us to look at the relative improvements in prediction performance, rather than claiming superiority of one method over another based on a single, arbitrary metric.

A preliminary observation is that the dataset selected is non-trivial for chemogenomics modeling challenges. Indeed, the baseline family QSAR model, which assumed all ligands to exhibit bioactivity on a single "average" and fictious GPCR failed to reach the accuracy level of both single endpoint models and of CG approaches. This ensures that herein the reported models better encapsulate specific knowledge of compound–protein interactions than could be expected from the average likelihood of a ligand to be a GPCR-binding molecule. Nonetheless, the family QSAR model is statistically sound, highlighting once more that such approaches could be useful for generic target-family directed library design.

The first and clearly validated benefit of chemogenomic modeling stems from inductive transfer. An IT implementation in CG simply requires assignment of each CPI entry to the concerned target, which in this work practically translates to concatenation of target identity fingerprints to ligand descriptors. There is no requirement for structural or functional information about the target set. IT can also serve to enhance single-point QSAR models, by using predictive models of other targets as a complement of information to the molecular descriptors of the ligand. Interestingly, all of the IT-driven approaches, single- or multi-endpoint, are very similar in terms of cross-validated prediction propensities, and form a homogeneous cluster in strategy space.

Any potential benefits beyond IT enhancement should arise from explicit learning based on genuine target information. However, the simple act of concatenating genuine protein descriptors to ligand descriptors does not yet imply that explicit learning will occur. Indeed, the algorithm could simply exploit these descriptors as non-empty target labels and operate analogously to above. However, explicit learning is an absolute prerequisite for rational predictive extrapolation to orphan targets. In the absence of EL, such an extrapolation will be as serendipitous as the simple hypothesis that the deorphanization could be done by a substitute model trained on a related endpoint. Relevant protein descriptors are a prerequisite to EL. This study explored both classical protein sequence-based terms as default descriptors as well as a protein description profile built from experimentally observed ligation pattern similarities.

The results of the cross-validation prediction challenge failed to evidence any significant performance between IT-enhanced and EL-enabled CG approaches. However, this observation alone is insufficient evidence to assess the actual role played by EL in the enabled models. Indeed, in a cross-validation scenario, IT is fully competent to adjust ligand feature weights even for data-poor endpoints. The only test that can distinguish between genuine EL enhancement and IT-driven benefits is deorphanization.

Deorphanization challenges showed that serendipitous deorphanization by substitution with a single-endpoint model of a target analog was often observed within the datasets used. This was in spite of removal of 1/3 of the training targets which became presumed orphans, a more challenging scenario than the typical "leave-one-target-out" formulation. IT-enhanced CG models, being as ignorant of target nature, are not per se better suited for deorphanization than single-endpoint approaches. They were indeed found to be occasionally successful in this challenge, albeit for targets different than those deorphanized by substitution. Deorphanization by the IT-enhanced CG approach is a consensus effect driven by multiple near analogs in the training set. This consensus is beneficial with respect to presumed orphans (alpha-1B or serotonin-1B) which had multiple, equally relevant analogs (alpha-1A/D or serotonin 1A/D). In contrast, targets that qualify for deorphanization by substitution were found to be those that had a dominant analog remaining in the training set.

EL-enabled models failed to exceed performances of IT-driven approaches. Strategy space clustering showed that the two approaches continued to cluster closely together as they had in cross-validation tests. This is unfortunately positive proof that herein studied EL-enabled models did not live up to our expectations. This raises questions about the usefulness of sequence-based protein descriptors in attempted EL-driven modeling. Furthermore, the injection of experimental ligation affinity patterns (the SIMFP descriptor) did not seem to significantly boost deorphanization performance either (Table 7). Or, SIMFP may be regarded as quasi-ideal protein descriptors, rendering the actual protein–protein similarity levels in terms of ligand recognition behavior, which represents a challenging objective to reproduce on the basis of so-far envisaged protein descriptors, including 3D terms [47, 48]. Therefore, the "window of opportunity" for useful EL-enhanced approaches may be quite narrow, because on one hand simple sequence-based protein descriptors do not seem to be sufficiently information-rich, and on the other hand, computationally expensive site-based 3D descriptors are not only limited in use to well-known non-orphan targets, but also should not be more compute-intensive than structure-based approaches. At this point, it appears that EL-enabled models are indeed indistinguishable, in all respects, from pure IT-driven approaches.

Based on these observations and insights, several perspectives merit discussion. A first would be the need for research into protein descriptors in view of development of active-site aware terms that might boost explicit learning. Additionally there is a need to explore different machine learning techniques that explicitly support a formulation of ligand feature weights as functions of protein description. Notably, an enforcement of training with emphasis placed on protein information may prove useful.

Finally, this paper highlights the importance of strict and extensive benchmarking protocols, including baseline experiments such as family QSAR construction, and reporting of deorphanization by substitution. It also makes use of the strategy space concept to avoid excessive optimization over any individually preferred evaluation metric. No longer do models simply receive a ranking based on an absolute metric value, but rather the *relative* performance through a joint description of metrics is used to provide a more robust assessment of results.

We suggest that future work in chemogenomics modeling continue to incorporate the concept of relative performance for reporting results. This is important, as many successful applications of CG are difficult to interpret, in absence of such in-depth benchmarking. For example, a recent study [49] reports a surprising lack of sensitivity of CG performance with respect to the choice of protein descriptors. This observation might prove supportive of the herein reported conclusions. The hypothesis that, like in this study, the role of the 13 therein used sets of protein descriptors did not go beyond acting as ”target labels” and hence enabling IT—but not EL—appears in a very favorable light, but is not demonstrated.

Also, the wealth of positive feedback from CG approaches is, per se, not in contradiction with the herein reported results.

Supplementary Information

Smiles-activity data (columns 1 and 2 of ref.smi_act_info) and treePH03 descriptors (others available upon request) for ligands associated to each target *T* (individual directories) are provided in a tar archive, together with a dictionary file relating the target T code (directory label) to actual identifiers (name, ChEMBL ID). Experimental (col. 2) *vs.* predicted results (col. 3) for the ligands of each target (col. 1) are reported together with the standard deviation of prediction (col. 4) in the ’allpred’ files, named by the corresponding strategy. ’d’-prefixed files are deorphanization challenge results, the other are cross-validated predictions at model building stage.

Acknowledgments The authors thank the High Performance Computing centers of the Universities of Strasbourg, France, and Cluj, Romania, for having hosted a part of the herein reported calculations. This work was also supported in part by a Grant-in-Aid for Young Scientists from the Japanese Society for the Promotion of Science [Kakenhi (B) 25870336]. Special thanks are given to Professor Jürgen Bajorath of the University of Bonn for extracting and cleaning the herein used datasets. This research was additionally supported by the Funding Program for Next Generation World-Leading Researchers as well as the CREST program of the Japan Science and Technology Agency. J. B. Brown and Yasushi Okuno are supported in part by grants from Chugai Pharmaceutical Co. Ltd. and Mitsui Knowledge Co. Ltd.

References

1. Abernethy J, Bach F, Evgeniou T, Vert JP (2009) A new approach to collaborative filtering: operator estimation with spectral regularization. J Mach Learn Res 10:803–826
2. Argyriou A, Evgeniou T, Pontil M (2008) Convex multi-task feature learning. Mach Learn 73(3):243–272
3. Bock JR, Gough DA (2002) A new method to estimate ligand-receptor energetics. Mol Cell Proteomics 1(11):904–910
4. Bock JR, Gough DA (2005) Virtual screen for ligands of orphan G protein-coupled receptors. J Chem Inf Model 45(5):1402–1414
5. Bonachera F, Horvath D (2008) Fuzzy tricentric pharmacophore fingerprints. 2. Application of topological fuzzy pharmacophore triplets in quantitative structure–activity relationships. J Chem Inf Model 48(2):409–425
6. Bonachera F, Parent B, Barbosa F, Froloff N, Horvath D (2006) Fuzzy tricentric pharmacophore fingerprints. 1—topological fuzzy pharmacophore triplets and adapted molecular similarity scoring schemes. J Chem Inf Model 46:2457–2477
7. Brown J, Nijima S, Okuno Y (2013) Compound–protein interaction prediction within chemogenomics: theoretical concepts, practical usage, and future directions. Mol Inf 32:906–921
8. Brown J, Okuno Y (2012) Systems biology and systems chemistry: new directions for drug discovery. Chem Biol 19(1):23–28
9. Caruana R (1997) Multitask learning. Mach Learn 28(1):41–75
10. Chang CC, Lin CJ (2011) LIBSVM: a library for support vector machines. ACM Trans Intell Syst Technol 2(27):1–27
11. Collantes E, Dunn W (1995) Amino acid side chain descriptors for quantitative structure–activity relationship studies of peptide analogs. J Med Chem 38(14):2705–2713
12. Evgeniou T, Micchelli CA, Pontil M (2005) Learning multiple tasks with kernel methods. J Mach Learn Res 6:615–637
13. Frimurer T, Ulven T, Elling C, Gerlach LO, Kostenis E, Hogberg T (2005) A physicogenetic method to assign ligand–binding relationships between 7TM receptors. Bioorg Med Chem Lett 15:3707–3712
14. Gaulton A, Bellis LJ, Bento AP, Chambers J, Davies M, Hersey A, Light Y, McGlinchey S, Michalovich D, Al-Lazikani B, O-verington JP (2011) Chembl: a large-scale bioactivity database for drug discovery. Nucl Acids Res 40(D1):D1100–D1107
15. Gozalbes R, Rolland C, Nicola E, Paugam MF, Coussy L, Horvath D, Barbosa F, Mao B, Revah F, Froloff N (2005) QSAR strategy and experimental validation for the development of a GPCR focused library. QSAR Comb Sci 24(4):508–516
16. Harrell F (2001) Regression modeling strategies: with applications to linear models, logistic regression, and survival analysis. Graduate texts in mathematics. Springer, Berlin
17. Horvath D, Bonachera F, Solov'ev V, Gaudin C, Varnek A (2007) Stochastic versus stepwise strategies for quantitative structure–activity relationship generation—how much effort may

the mining for successful QSAR models take? J Chem Inf Model 47:927–939

18. Horvath D, Marcou G, Varnek A (2013) Do not hesitate to use tversky—and other hints for successful active analogue searches with feature count descriptors. J Chem Inf Model 53(7):1543–1562
19. Hurle MR, Yang L, Xie Q, Rajpal DK, Sanseau P, Agarwal P (2013) Computational drug repositioning: from data to therapeutics. Clin Pharmacol Ther 93(4):335–341
20. Ivanciuc O (2007) Applications of support vector machines in chemistry. Wiley, New York, pp 291–400
21. Jacob L, Hoffmann B, Stoven V, Vert JP (2008) Virtual screening of GPCRS: an in silico chemogenomics approach. BMC Bioinform 9(1):363
22. Jacob L, Vert JP (2008) Protein–ligand interaction prediction: an improved chemogenomics approach. Bioinformatics 24(19):2149–2156
23. Kontijevskis A, Komorowski J, Wikberg JES (2008) Generalized proteochemometric model of multiple cytochrome p450 enzymes and their inhibitors. J Chem Inf Model 48(9):1840–1850
24. Kontijevskis A, Prusis P, Petrovska R, Yahorava S, Mutulis F, Mutule I, Komorowski J, Wikberg J (2007) A look inside HIV resistance through retroviral protease interaction maps. PLoS Comput Biol 3:e48
25. Lapins M, Eklund M, Spjuth O, Prusis P, Wikberg J (2008) Proteochemometric modeling of hiv protease susceptibility. BMC Bioinform 9(1):181
26. Lapinsh M, Prusis P, Gutcaits A, Lundstedt T, Wikberg J (2001) Development of proteo-chemometrics: a novel technology for the analysis of drug–receptor interactions. Biochim Biophys Acta 1525:180–190
27. Leslie CS, Eskin E, Cohen A, Weston J, Noble WS (2004) Mismatch string kernels for discriminative protein classification. Bioinformatics 20(4):467–476
28. Li S, Xi L, Wang C, Li J, Lei B, Liu H, Yao X (2009) A novel method for protein–ligand binding affinity prediction and the related descriptors exploration. J Comput Chem 30(6):900–909
29. Li ZR, Lin HH, Han LY, Jiang L, Chen X, Chen YZ (2006) PROFEAT: a web server for computing structural and physicochemical features of proteins and peptides from amino acid sequence. Nucl Acids Res 34(Suppl. 2):W32–W37
30. Medina-Franco JL, Giulianotti MA, Welmaker GS, Houghten RA (2013) Shifting from the single to the multitarget paradigm in drug discovery. Drug Discov Today 18(9–10):495–501
31. Mikhalev AA, Shpilrain V, Yu JT (2004) The embedding problem. In: Borwein P, Borwein J (eds) Combinatorial methods. CMS books in mathematics. Springer, New York, pp 108–128
32. Pelikan M, Goldberg DE, Lobo FG (2002) A survey of optimization by building and using probabilistic models. Comput Optim Appl 21:5–20
33. Rao HB, Zhu F, Yang GB, Li ZR, Chen YZ (2011) Update of PROFEAT: a web server for computing structural and physicochemical features of proteins and peptides from amino acid sequence. Nucl Acids Res 39(Suppl. 2):W385–W390
34. Rosenbaum L, Dorr A, Bauer MR, Boeckler FM, Zell A (2013) Inferring multi-target QSAR models with taxonomy-based multitask learning. J Cheminform 5:1–20
35. Ruggiu F, Gizzi P, Galzi JL, Hibert M, Haiech J, Baskin I, Horvath D, Marcou G, Varnek A (2014) Quantitative structure–property relationship modeling: a valuable support in high-throughput screening quality control. Anal Chem 86(5):2510–2520
36. Ruggiu F, Marcou G, Varnek A, Horvath D (2010) Isida property-labelled fragment descriptors. Mol Inform 29(12):855–868
37. Sandberg M, Eriksson L, Jonsson J, Sjostrom M, Wold S (1998) New chemical descriptors relevant for the design of biologically active peptides. A multivariate characterization of 87 amino acids. J Med Chem 41:2481–2491
38. Schölkopf B, Tsuda K, Vert J (2004) Kernel methods in computational biology. MIT, Boston, MA, USA
39. Smola AJ, Schlkopf B (2004) A tutorial on support vector regression. Stat Comput 14(3):199–222
40. Strombergsson H, Daniluk P, Kryshtafovych A, Fidelis K, Wikberg J, Kleywegt G, Hvidsten T (2008) Interaction model based on local protein substructures generalizes to the entire structural enzyme–ligand space. J Chem Inf Model 48:2278–2288
41. Tetko IV (2002) Neural network studies. 4. Introduction to associative neural networks. J Chem Inf Comput Sci 42(3):717–728
42. Van Westen G, Wegner J, Geluykens P, Kwanten L, Vereycken I, Peeters A, IJzerman A, Van Vlijmen H, Bender A (2011) Which compound to select in lead optimization? Prospectively validated proteochemometric models guide preclinical development. PLoS One 6:e27518
43. Van Westen G, Wegner J, Ijzerman A, Van Vlijmen H, Bender A (2011) Proteochemometric modeling as a tool for designing selective compounds and extrapolating to novel targets. Med Chem Commun 2:16–30
44. Varnek A, Gaudin C, Marcou G, Baskin I, Pandey AK, Tetko IV (2009) Inductive transfer of knowledge: application of multi-task learning and feature net approaches to model tissue-air partition coefficients. J Chem Inf Model 49(1):133–144
45. Varnek A, Tropsha A (2009) Chemoinformatics: approaches to virtual screening. Royal Society of Chemistry. Cambridge, USA
46. Wassermann AM, Geppert H, Bajorath J (2009) Ligand prediction for orphan targets using support vector machines and various target-ligand kernels is dominated by nearest neighbor effects. J Chem Inf Model 49(10):2155–2167
47. Weill N, Rognan D (2009) Development and validation of a novel protein–ligand fingerprint to mine chemogenomic space: application to G protein-coupled receptors and their ligands. J Chem Inf Model 49(4):1049–1062
48. Weill N, Rognan D (2010) Alignment-free ultra-high-throughput comparison of druggable proteinligand binding sites. J Chem Inf Model 50(1):123–135
49. van Westen G, Swier R, Cortes-Ciriano I, Wegner J, Overington J, IJzerman A, Van Vlijmen H, Bender A (2013) Benchmarking of protein descriptors in proteochemometric modeling (part 2): modeling performance of 13 amino acid descriptors. J Cheminform 5:42
50. van Westen GJP, Wegner JK, Ijzerman AP, van Vlijmen HWT, Bender A (2010) Proteochemometric modeling as a tool to design selective compounds and for extrapolating to novel targets. MedChemComm 2(1):16–30
51. Yabuuchi H, Niijima S, Takematsu H, Ida T, Hirokawa T, Hara T, Ogawa T, Minowa Y, Tsujimoto G, Okuno Y (2011) Analysis of multiple compound–protein interactions reveals novel bioactive molecules. Mol Syst Biol 7(472)