J Comput Aided Mol Des (2014) 28:509–547
DOI 10.1007/s10822-014-9740-4

The use of docking-based comparative intermolecular contacts analysis to identify optimal docking conditions within glucokinase and to discover of new GK activators

Mutasem O. Taha · Maha Habash · Mohammad A. Khanfar

Received: 14 November 2013 / Accepted: 25 February 2014 / Published online: 8 March 2014
© Springer International Publishing Switzerland 2014

Abstract Glucokinase (GK) is involved in normal glucose homeostasis and therefore it is a valid target for drug design and discovery efforts. GK activators (GKAs) have excellent potential as treatments of hyperglycemia and diabetes. The combined recent interest in GKAs, together with docking limitations and shortages of docking validation methods prompted us to use our new 3D-QSAR analysis, namely, docking-based comparative intermolecular contacts analysis (dbCICA), to validate docking configurations performed on a group of GKAs within GK binding site. dbCICA assesses the consistency of docking by assessing the correlation between ligands' affinities and their contacts with binding site spots. Optimal dbCICA models were validated by receiver operating characteristic curve analysis and comparative molecular field analysis. dbCICA models were also converted into valid pharmacophores that were used as search queries to mine 3D structural databases for new GKAs. The search yielded several potent bioactivators that experimentally increased GK bioactivity up to 7.5-folds at 10 μM.

Keywords Docking · LigandFit · Glucokinase enzyme · dbCICA · In-silico screening · In vitro testing

Electronic supplementary material The online version of this article (doi:10.1007/s10822-014-9740-4) contains supplementary material, which is available to authorized users.

M. O. Taha (✉) · M. A. Khanfar
Drug Discovery Unit, Department of Pharmaceutical Sciences, Faculty of Pharmacy, The University of Jordan, Amman, Jordan
e-mail: mutasem@ju.edu.jo

M. Habash
Department of Pharmaceutical Chemistry and Pharmacognosy, Faculty of Pharmacy, Applied Science University, Amman, Jordan

Introduction

Glucokinase (GK), also referred to as hexokinase IV or D, is a member of the hexokinases family. It is predominantly expressed in the liver and pancreas. GK catalyses the phosphorylation of glucose to glucose-6-phosphate (G6P) via adenosine triphosphate (ATP) and Mg^{2+}. GK exerts high control in hepatic glucose metabolism: It acts as key player in the fed state by influencing glucose uptake, while in the fasted state it controls glucose production [1].

GK has a unique kinetic profile compared to other hexokinases: It has low affinity to glucose at low glucose concentrations; however, it becomes significantly more active at higher glucose levels. This sigmoidal response to glucose concentration is referred to as 'positive kinetic cooperativity for glucose' and it seems to be related to the unique kinetic transition forms of GK [2, 3].

The combination of positive kinetic cooperativity, low affinity to glucose at low glucose concentrations, and lack of end-product inhibition make GK activators (GKAs) of excellent potential as treatments for hyperglycemia and diabetes [4]. Activating hepatic GK activity results in enhanced glucose utilization in response to hyperglycemia, while in the pancreas, increased GK activity results in enhanced insulin secretion. Therefore, activation of GK should result in better control of blood glucose levels via both hepatic and pancreatic pathways. Additionally, reduction in GK activity in response to low glucose levels reduces the possibility of hypoglycaemia during treatment with GKAs [5].

Co-crystallization of GKAs within GK shows that these compounds bind to an allosteric pocket [2]. GKAs increase the affinity of GK for glucose by stabilizing certain 'closed' conformation of the kinase [2, 3].

The main focus of recent efforts towards the development of new GKAs concentrate on structure-based ligand

design efforts [6–8]. To date, 11 human GKa X-ray complexes are documented in the Protein Data Bank (e.g., PDB codes: 3ID8, 3IDH, 3FGU, 3H1V, 3IMX, 3AOI, 3GOI, 3FRO, 3F9M, 1V4S and 1V4T [9–16] of resolution range from 1.50 to 3.40 Å.

However, despite the limitations of crystallographic structures [17–20, 25], structure-based drug design is still considered one of the most important tools in drug discovery [21–23] including GKAs [9–16]. Docking, a significant element of structure-based design, involves virtual fitting of ligands into corresponding binding site(s) employing algorithms that rely on force fields to calculate attractive and repulsive interactions between complementary groups within virtual ligand–protein complexes [21–25].

Molecular docking is essentially a conformational sampling procedure in which various docked conformations are explored to identify the best one. However, docking can be a very challenging problem given the degree of conformational flexibility at the ligand-macromolecular level [26–28]. Docking programs employ diverse methodologies to evaluate different ligand conformations within binding pockets [29–40]. However, conformational sampling must be guided by a scoring function to evaluate the fitness between the protein and the ligand [24, 41–46]. The final docked conformations are also selected according to their scores. The accuracy of the scoring function has a major impact on the quality of molecular docking results [42, 48, 49]. Scoring functions can be roughly grouped into three categories: force field methods [31, 32, 35, 50]; empirical scoring functions [30, 51–55], and knowledge-based potentials [56–60].

However, the underlying molecular interactions in ligand-receptor binding are highly complex and various terms should be considered to quantify the free energy of binding [17, 47, 61, 63–65]. Incidentally, although certain modifications were recently proposed to improve the outcomes of scoring functions (e.g., ligand efficiency indices) [66], the sheer complexity of ligand-receptor interactions could still wane the ability of scoring functions to rank different potential ligand–receptor complexes [17, 21, 23, 47, 61, 62]. Accordingly, the molecular modeler must find the optimal combination of docking/scoring algorithms capable of correct ranking of docked conformers/poses for potential ligands within a certain binding pocket.

Furthermore, the molecular modeler must decide whether to leave or remove crystallographically explicit water molecules in the binding site prior to ligand docking [67–74]. Although recent functions were proposed to identify those bound water molecules that should be included in docking algorithms [72–74], the hydrated binding site may be one of the many structure conformations available to the receptor, and different ligands will have a different ability to select either hydrated or non-hydrated receptor binding site conformations [72–74]. Moreover, the fact that crystallographic structures lack information on hydrogen atoms means that it should be appropriately assumed, prior to docking, whether the ligand's ionizable moieties embedded within the binding site exist in their ionized form or not [71]. Reliance on pKa values can be misleading since ligand ionizabilities depend on their local microenvironments within the binding pockets [71]. Additionally, the proper treatment of molecules that can tautomerize is challenging. Docking must involve the decision as to which tautomers to include in the docking and how to account for tautomerization in the scoring [75].

These problems make it vital to validate docked conformers/poses for subsequent structure-based discovery or design [76]. Current docking validation methods can be classified into: (i) Self-docking [77]: co-crystallized ligands are removed from their corresponding binding sites and re-docked employing the docking configuration under evaluation. However, this approach overlooks the fact that docking experiments are usually performed to dock ligands into binding pockets imprinted by other co-crystallized ligand(s) [77]. (ii) Testing the ability of particular docking configuration to classify compounds in structural databases into actives and inactive [78–80]. Nevertheless, this approach suffers from a major drawback: it assumes inactivity of decoy molecules despite lack of supporting evidence [81]. (iii) Validation through 3D-QSAR methods. In this case, a particular docking configuration is considered valid if it succeeds in aligning a set of known ligands (i.e., into the binding pocket) in a 3D alignment capable of explaining bioactivity variation, e.g., via CoMFA [64, 82, 83]. However, this approach is quite time-consuming and laborious.

The combined recent interest in GKAs, together with docking pitfalls and shortages of docking validation methods prompted us to use our new 3D-QSAR analysis to validate docking configurations of GKAs within GK, namely, docking-based comparative intermolecular contacts analysis (dbCICA) [84, 85]. This approach is based on the number and quality of contacts between docked ligands and amino acid residues within the particular binding pocket. dbCICA assesses a docking configuration based on its ability to align a set of ligands (i.e., within a corresponding binding pocket) in such a way that potent ligands come into contact with binding site spots distinct from those approached by low-affinity ligands and vice versa. In other words, dbCICA evaluates the consistency of docking by assessing the correlation between ligands' affinities and their contacts with binding site spots [84, 85].

We implemented dbCICA to evaluate a variety of docking-scoring conditions to identify optimal docking parameters capable of aligning a group of GKAs within a

GK binding site in such a way to be consistent with their bioactivities. A list of 71 GKAs (Table 1) were docked into the binding pocket of GK employing LigandFit docking engine via 6 scoring functions. The docked compounds were split into two subgroups: the first set is unionizable (i.e., within physiological pH, **1-41**, Table 1) and were docked as unionized into the binding site in the presence and absence of explicit hydration water molecules (hydrous and anhydrous binding site). The second set of GKAs is ionizable (**42-71**, Table 1) and was docked into GK in two ionization states (ionized and unionized) and two binding site hydration states (hydrous and anhydrous).

Subsequently, dbCICA modeling was employed in both cases as gauge to measure the success of each set of docking parameters. Thereafter, optimal dbCICA models were employed as templates to generate pharmacophore models that were employed as search queries to screen the National Cancer Institute (NCI) structural database for new GKAs. Hits were subsequently bioassayed. Several hits illustrated interesting biological properties.

Materials and methods

Molecular modeling

Case one: unionized GK activators data set

The structures of 41 GKAs were assembled from published literature (**1–41**, Table 1) [86, 87]. They were carefully collected such that they were bioassayed employing similar conditions in order to allow proper QSAR correlation (dbCICA analysis). The remaining compounds in Table 1 (**42–71**) were used also for addition of steric constrains to the successful models obtained from first activators set (using HipHopRefine, see "Addition of exclusion volumes" section under Experimental).

The in vitro bioactivities of the collected activators were expressed as the concentration of the test compound that activated GK enzyme by 50 % i.e. EC_{50}. Table 1 shows the structures and EC_{50} values of the collected activators. The logarithm of measured EC_{50} (μM) values were used in QSAR analysis, thus correlating the data linear to the free energy change. In cases where EC_{50} values were expressed as being "not active", i.e., **7, 11, 16** and **17**, we assumed they exhibited EC_{50} values of 3,000 μM. These assumptions are necessary to allow statistical correlation and QSAR analysis. The logarithmic transformation of EC_{50} values should minimize any potential error resulting from this assumption as well as from minor discrepancies in bioassay conditions among different authors.

The two dimensional structures of collected GKAs were sketched in ChemDraw Ultra (Version 11.0). The structures were subsequently converted into reasonable three-dimensional representations employing the rule-based methods implemented in DS 2.0 and were saved in SD format for subsequent experiments.

Case two: ionizable GK activators data set

The structures of 30 GKAs were assembled from published literature (**42–71**, Table 1) [88, 89]. They were all bioassayed employing similar conditions. The in vitro bioactivities of the collected activators were expressed as the concentration of the test compound that activated GK enzyme by 50 % (i.e., EC_{50}). The logarithm of measured EC_{50} (μM) values were used in QSAR analysis. In cases where EC_{50} values were expressed as being >10 μM; i.e., **44–47**, **49** and **51**, we assumed $EC_{50} = 200$ μM (we aim by assigning this arbitrary value to denote the difference from totally inactive compounds in the first case). This assumption is necessary to allow statistical correlation and QSAR analysis. Again logarithmic transformation should minimize any potential error resulting from this assumption as well as from minor discrepancies in bioassay conditions among different authors.

Preparation of GK crystal structure

The 3D coordinates of GK were retrieved from the Protein Data Bank (PDB code: 1V4S, resolution: 2.3 Å). Hydrogen atoms were added to the protein utilizing DS 2.0 templates for protein residues. Gasteiger-Marsili charges were assigned to the protein atoms as implemented within DS 2.0 [90].

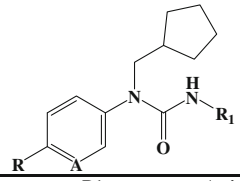
The protein structure was utilized in subsequent docking experiments without energy minimization. Explicit water molecules were either kept or removed according to the required docking conditions, i.e., docking in the presence or absence of explicit water molecules.

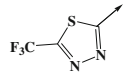
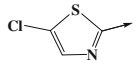
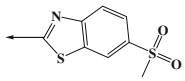
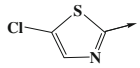
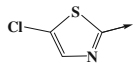
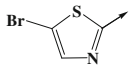
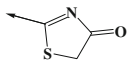
LigandFit docking and scoring

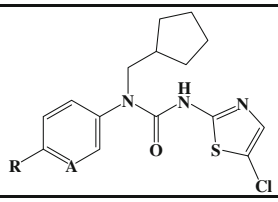
Docking experiments were conducted employing LigandFit docking engine. LigandFit considers the flexibility of the ligand and treats the receptor as rigid [91–93]. Implemented docking configurations and their theoretical explanations are shown in details in section (SM-2) in the supplementary material. High ranking docked conformers/poses were scored using 6 scoring functions: Jain [50], Ligscore1, Ligscore2 [43, 91], PLP1 [92], PLP2 [54] and PMF [56–58].

Considering each scoring function in turn, the highest scoring docked conformer/pose was selected for each activator for subsequent 3D-QSAR modeling. This, for

Table 1 Compounds involved in GK structure based modeling



Compound No	A	R	R1	Activity EC_{50} (μM)	Ref
1	**CH**	SO_2Me		16.0	[84]
2	**CH**	SO_2Me		34.2	*ibid*
3	**CH**	SO_2Me		1.5	*ibid*
4[a]	**CH**	SO_2Me		32.9	*ibid*
5[a]	**CF**	F		16.3	*ibid*
6	**CF**	F		2.6	*ibid*
7	**CF**	F		NA	*ibid*
8	**N**	OMe		19.2	*ibid*
9[a]	**N**	OMe		2.7	*ibid*
10	**N**	OMe		2.0	*ibid*
11	**N**	OMe		NA	*ibid*
12	**N**	OMe		29.1	*ibid*



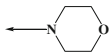
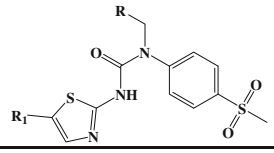
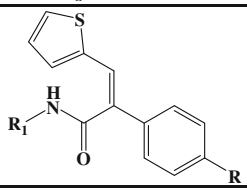
Compound No	A	R	Activity EC 50 (μM)	Ref
13[a]	CH	SO_2NH_2	1.8	*ibid*
14	CH	CF_3	1.3	*ibid*
15	CH	CN	2.4	*ibid*
16	CH		NA	*ibid*
17[a]	CH	OPh	NA	*ibid*
18[a]	CCl	Me	0.4	*ibid*
19	CF	Br	0.7	*ibid*
20	C(OMe)	H	1.4	*ibid*
21	$C(NHCOCH_3)$	H	0.3	*ibid*
22[a]	$C(CONH_2)$	H	1.0	*ibid*
23	N	CN	7.7	*ibid*

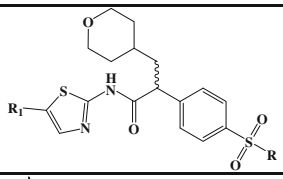
Table 1 continued



Compound No	R_1	R	Activity EC 50 (μM)	Ref
24	Cl		6.6	*ibid*
25[a]	Cl		6.8	*ibid*
26	Br		6.5	*ibid*
27	Br		24.1	*ibid*
28	Cl		11.5	*ibid*
29	Cl		21.6	*ibid*

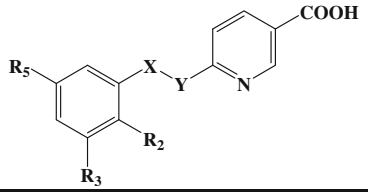


Compound No	R_1	R	Activity EC 50 (μM)	Ref
30		H	26.01	[85]
31		H	10.77	*Ibid*
32		SO_2Me	0.56	*ibid*

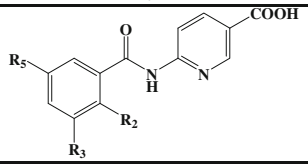


Compound No	R_1		R	Activity EC 50 (μM)	Ref
33	H		Me	2.30	*ibid*
34	H		Me	3.47	*ibid*
35	H		Me	1000	*ibid*
36	H			0.57	*ibid*
37	Me			0.58	*ibid*
38	F			0.13	*ibid*
39	F		Me	1.43	*ibid*
40	F			0.14	*ibid*
41	F			0.07	*ibid*

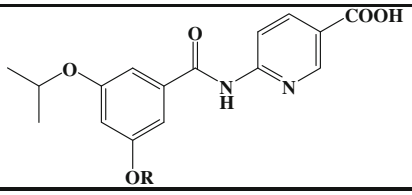
Table 1 continued



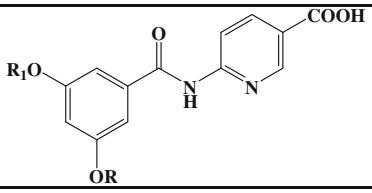
Compound	R2	R3	R5	X-Y	EC50 (μM)	Ref
42[b]	-OCH_2Ph	H	-SCH_3	-CH=CH-	3.20	[86]
43	H	-OCH_2Ph	-OCH_2Ph	-CH2CH2-	0.91	*ibid*
44	H	-OCH_3	-OCH_3	-OCH2-	>10	*ibid*
45	H	-OCH_2Ph	-OCH_2Ph	-CH2O-	>10	*ibid*
46	H	-OCH_2Ph	-OCH_2Ph	-NHCO-	>10	*ibid*
47[a]	-OCH_2Ph	H	-SCH_3	-NHCO-	>10	*ibid*
48	-OCH_2Ph	H	-SCH_3	-CONH-	4.22	*ibid*



Compound	R2	R3	R5	EC50 (μM)	Ref
49	-OCH_2Ph	H	H	>10	*ibid*
50	H	-OCH_2-*o*-Cl-Ph	H	0.91	*ibid*
51[a]	-OCH_2-*o*-Cl-Ph	-OCH_2-*o*-Cl-Ph	H	>10	*ibid*
52	H	-$OCH(CH_3)_2$	-$OCH_2CH(CH_3)_2$	0.57	*ibid*
53	H	-OCH_2Ph	-OCH_2Ph	0.40	*ibid*
54	H	-OCH_2-*o*-F-Ph	-OCH_2-*o*-F-Ph	0.09	*ibid*



Compound	R	EC50 (μM)	Ref
55	-CH_2CH_2-4-THP	1.33	*ibid*
56[a]	-CH_2-*C*Pent	0.65	*ibid*
57	-CH_2CH_2- *C*Pent	0.17	*ibid*
58	-CH_2CH_2-3-pyridyl	1.26	*ibid*
59[a]	-CH_2CH_2-4-pyridyl	1.78	*ibid*
60	-CH_2CH_2Ph	0.13	*ibid*
61	-CH_2CH_2-3-thiophene	0.09	*ibid*
62[a]	-CH_2Ph	0.29	*ibid*
63	-$CH_2CH_2CH_2Ph$	0.42	*ibid*
64	-CH_2-*O*-F-Ph	0.10	*ibid*



Compound	R	R1	EC50 (μM)	Ref[a]
65	*(S)*- $CH(CH_3)Ph$	$CH(CH_3)_2$	0.11	[87]
66	*(R)*- $CH(CH_3)Ph$	$CH(CH_3)_2$	0.95	*ibid*
67	*(S)*- $CH(CH_3)CH_2OCH_3$	$CH(CH_3)_2$	0.61	*ibid*
68	*(R)*- $CH(CH_3)CH_2OCH_3$	$CH(CH_3)_2$	5.51	*ibid*
69	*(S)*- $CH(CH_3)CH_2Ph$	$CH(CH_3)_2$	0.02	*ibid*
70[a]	*(R)*- $CH(CH_3)CH_2Ph$	$CH(CH_3)_2$	0.09	*ibid*
71	*(S)*- $CH(CH_3)CH_2Ph$	*(S)*-$CH(CH_3)CH_2OCH_3$	0.03	*ibid*

[a] These compounds were employed as the external test subset in QSAR and CoMFA modeling

[b] Compounds (**42–71**) were used for steric refinement of dbCICA generated Hypo1, Hypo2 and Hypo3

Table 2 Different liganfit-based docking conditions for compounds (**1–41**, Table 1), their corresponding best dbCICA parameters and statistical criteria

Docking conditions			Optimal dbCICA parameters			dbCICA statistical criteria		
Ligands' ionization state	Explicit water[a]	Scoring functions	Contacts distance threshold (Å)[b]	Number of positive contacts[c]	Number of negative contacts[d]	r^2_{30} [e]	r^2_{LOO} [f]	r^2_{5-fold} [g]
Unionized	Present	Jain	2.5	9	10	0.59	0.55	0.56
		Ligscore 1	2.5	7	5	0.58	0.53	0.54
		Ligscore 2	2.5	7	10	0.62	0.57	0.59
		PLP1	3.5	7	5	0.50	0.45	0.47
		PLP2	3.5	7	5	0.58	0.54	0.54
		PMF	2.5	6	10	0.63	0.59	0.58
	Absent	Jain	2.5	10	10	0.63	0.60	0.58
		Ligscore 1[h]	*3.5*	*10*	*10*	*0.78*	*0.75*	*0.76*
		Ligscore 2	3.5	10	5	0.66	0.63	0.63
		PLP1	3.5	7	10	0.69	0.66	0.64
		PLP2[h]	*2.5*	*7*	*10*	*0.67*	*0.64*	*0.67*
		PMF[h]	*2.5*	*4*	*10*	*0.68*	*0.65*	*0.67*

[a] Crystalographically explicit water of hydration

[b] Distance thresholds used to define ligand-binding site contacts

[c] Optimal number of combined (i.e., summed) bioactivity-enhancing ligand/binding site contacts

[d] Optimal number of combined (i.e., summed) bioactivity-disfavoring ligand/binding site contacts

[e] Non-cross-validated correlation coefficient for 41 training compounds

[f] Cross-validation correlation coefficients determined by the leave-one-out technique

[g] Cross-validation correlation coefficients determined by the leave-20 %-out technique repeated 5 times

[h] Italic parameters correspond to the best docking/scoring combinations in LigandFit-based docking

Table 3 Highest ranking dbCICA models for compounds (**1–41**, Table 1), their corresponding parameters and statistical criteria

dbCICA model	Ligands' ionization state	Explicit water[a]	Scoring function	Contacts distance threshold[b]	Number of positive contacts[c]	Number of negative contacts[d]	r^2_{30} [e]	r^2_{LOO} [f]	r^2_{5-fold} [g]	F-statistic
A-I	Unionized	Absent	Ligscore1	3.5	10	10	0.78	0.75	0.76	134.6
A-II	Unionized	Absent	PLP2	2.5	7	10	0.67	0.64	0.67	80.8
A-III	Unionized	Absent	PMF	2.5	4	10	0.68	0.65	0.67	82.3

The presented information in this table are extracted from Table 2 except for F-statistic, which is calculated from the correlation connecting $-\log(EC_{50})$ and the contacts sums of the corresponding dbCICA models

[a] Crystalographically explicit water of hydration

[b] Distance thresholds used to define ligand-binding site contacts

[c] Optimal number of combined (i.e., summed) bioactivity-enhancing ligand/binding site contacts

[d] Optimal number of combined (i.e., summed) bioactivity-disfavoring ligand/binding site contacts

[e] Non-cross-validated correlation coefficient for 41 training compounds

[f] Cross-validation correlation coefficients determined by the leave-one-out technique

[g] Cross-validation correlation coefficients determined by the leave-20 %-out technique repeated 5 times

example, resulted in six sets of 41 docked molecules (first case, Table 2) with scores corresponding to each scoring function. However, the docking and scoring cycle was repeated 2 times to cover the different combinations of docking conditions, i.e., explicit water molecules (hydrous and anhydrous binding site). Further details about scoring docked poses are shown in details in section (SM-2) in the supplementary material.

Docking-based comparative molecular contacts analysis (dbCICA)

The methodology of dbCICA can be described in the following sequential steps:

(i) High ranking docked pose/conformer of each ligand, based on a particular docking condition, is evaluated to identify its closest atomic neighbors in the binding pocket. Intermolecular atomic neighbors closer than (or equal to) certain predefined distance threshold are assigned an intermolecular contact value of "one", otherwise they are assigned a contact value of "zero". For example, if atom A in the docked ligand is positioned close to atom B in the binding pocket at a distance shorter that the predefined threshold, then this contact is assigned a value of 1. Distance assessment is performed automatically employing the Intermolecular Monitor implemented in *Discovery Studio version 2.5 (DS 2.5)* [93]. Eventually, this step yields a two-dimensional matrix with row labels corresponding to docked ligands (i.e., according to the particular docking-scoring configuration) and column labels corresponding to different binding site atoms. The matrix is filled with binary code, whereby "zeros" correspond to inter-atomic distances above the predefined threshold and "ones" for distances below (or equal) the predefined threshold. In the current study two distance thresholds were implemented:

Table 4 Critical binding site contact atoms proposed by optimal dbCICA models for compounds (**1–41**, Table 1)

dbCICA model[a]	Favored contact atoms (positive contacts)[b]		Disfavored contact atoms (negative contacts)[e]
	Amino acids and corresponding atom identities[c]	Weights[d]	
A-I	CYS252: HB1	2	ARG250:HD2; GLU221:CA; LE159:CD1; MET235:CG; SER64:N; VAL452:HG11; VAL452:HG21; VAL455:HG23; VAL62:CA; VAL62:HG11
	HIS218: HB2	2	
	ILE159: HG23	2	
	ILE211: CA	2	
	LYS459: NZ	2	
	MET210: CE	2	
	SER64: C	2	
	TYR61: O	3	
	TYR215: O[f]	2	
	VAL455: HB	1	
	VAL455: O	3	
A-II	ALA456: HB1	2	GLU221:HA; GLU221:HG1; GLU67:N; GLU67:OE2; ILE159:CG2; ILE159:HD12; MET235:CE; THR65:HN; VAL452:HA; VAL62:HG23
	ALA456: HB3	2	
	ILE211: HD11	2	
	LEU451: O	2	
	THR65: N	1	
	TYR214: CD1	1	
	ILE211: HA	2	
A-III	GLU221: HG2	3	ARG250:HD2; ARG397:HH21; GLN219:HA; GLU67:OE2; A:ILE159:HD12; VAL455:CG1; VAL455:CG2; VAL455:HB; VAL62:CA; VAL62:HG23
	MET235: HE3	1	
	TYR214: HE2	2	
	TYR215: HE2	1	

[a] As in Table 4

[b] Bioactivity-proportional ligand/binding site contacts

[c] Binding site amino acids and their significant atomic contacts Atom codes are as provided by the protein data bank file format (e.g., VAL455:O encodes for oxygen atom (O) of valine number 455) except for hydrogen atoms which were coded by DS 2.0

[d] Degree of significance (weight) of corresponding contact atom (see point III under "Genetic algorithm implementation in dbCICA modeling" section

[e] Bioactivity-disfavoring ligand/binding site contacts

[f] This feature emerged via dbCICA modeling of hydrogen-bond forming atoms

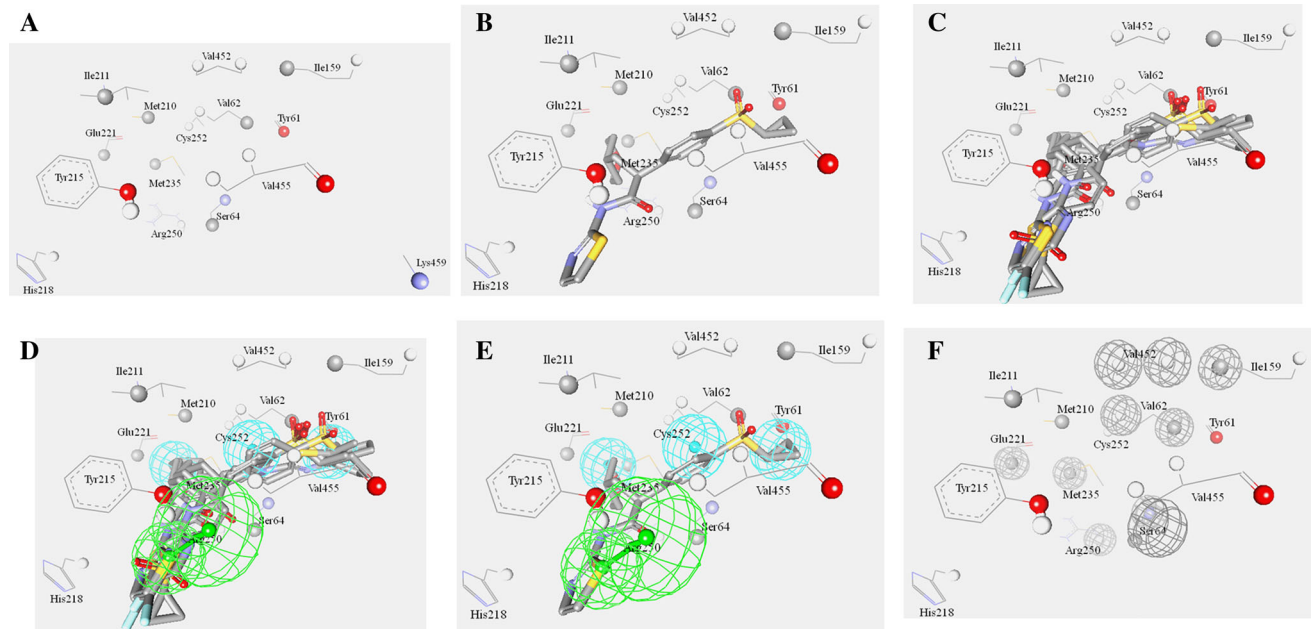


Fig. 1 Steps for manual generation of binding hypothesis HypoI as guided by dbCICA model A-I (Tables 3, 4): **a** the binding site moieties in dbCICA model A-I with significant contact atoms shown as *spheres*. **b** The docked pose of the well-behaved compound **36** ($EC_{50} = 0.57$ µM) within the binding pocket, **c** the docked poses of the well-behaved and potent compounds **36, 37, 38, 40** and **41**. **d** Manually placed pharmacophoric features onto chemical moieties common among docked well-behaved potent compounds **36, 37, 38, 40** and **41**. **e** The docked pose of **36** and how it relates to the proposed pharmacophoric features. **f** Exclusion spheres fitted against binding site atoms showing negative correlations with bioactivity (as emergent in dbCICA model A-I)

3.5 and 2.5. Therefore, two binary matrices (corresponding to each distance threshold) are constructed for each docking configuration (i.e., combination of docking engine, scoring function and binding site hydration status).

(ii) Each individual column in the matrix is regressed against the corresponding molecular bioactivities (i.e.,−log (EC_{50})). Columns that exhibit negative correlation with bioactivity are inverted, i.e., zeros are converted to ones and vice versa, and excluded from the subsequent step.

After excluding inverted columns (negative contacts or exclusion volumes), the resulting binary matrix (which is composed from positively correlated contact columns with bioactivity) is then subjected to genetic algorithm (GA)-based search for optimal summation of contacts columns capable of explaining bioactivity variation: In this step GA relies on the evolutionary operations of "crossover and mutation" to select optimal combination of columns that have their summation values collinear with bioactivity variation across training compounds (see GA parameters below). The best column-summation model (single model) is selected as representative dbCICA model.

GA can be instructed to output best dbCICA models resulting from any predefined number of ligand-receptor intermolecular positive contacts, e.g., best dbCICA models resulting from sets of 2 or 3 or 4 or 5, etc.… concomitant contacts (i.e., summed contact columns). In the current project we instructed GA to search for the best dbCICA models resulting from 2 contacts and repeat the scan to identify the best summation models for 3, 4, 5, 6, 7, 8, 9 and 10 contacts. Each set of summed contacts is treated independently to identify the corresponding dbCICA model in each case.

(iii) dbCICA algorithm has the option of allowing any particular positive contacts column to emerge up to three times in the optimal summation model, i.e., it allows variable weights for contacts. This is performed by implementing dual valued genes in the GA, in which every gene encodes for both the corresponding contacts column number and its weight. Column weights are initially randomly distributed in the first generation and subsequently subjected to mutation only (not cross-over) in GA. This option was allowed in the current project to identify intermolecular contacts of higher weights or contributions in the optimal dbCICA models.

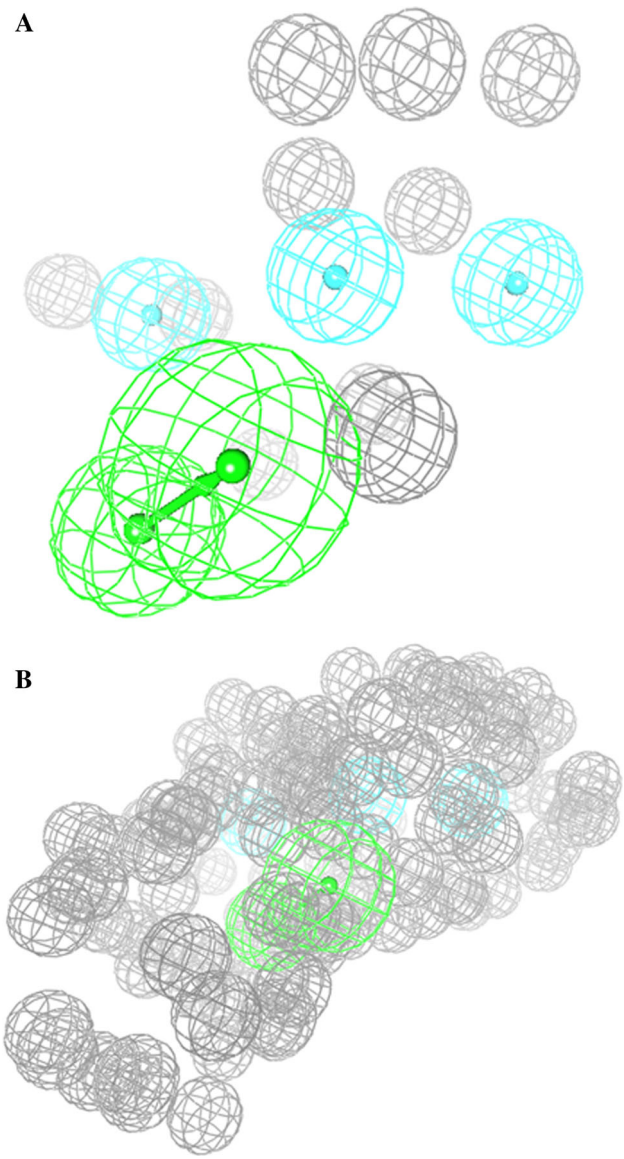


Fig. 2 **a** HypoA-I pharmacophoric features and exclusion spheres *light blue spheres* represent hydrophobic features, *vectored green spheres* represent hydrogen bond acceptor features, and *gray spheres* represent exclusion regions **b** Refined HypoA-I with 112 added exclusion spheres

(iv) After identifying optimal summation model(s) based on positive contacts (proportional to bioactivity), dbCICA implements GA to search for optimal summation model resulting from combining inverted columns (negatively proportional to bioactivity, see step ii) with the optimal positive summation model(s). The user has the option of choosing any number of negative contacts (excluded volumes or steric clashes) to emerge in the final dbCICA model. In the current project we implemented two exclusion settings; either five or ten negative contacts were allowed.

Genetic algorithm implementation in dbCICA modeling

The GA toolbox within MATLAB (Version R2007a) was adapted by implementing the following four basic components: the creation function, cross-over function, mutation function, and fitness function.

The creation function randomly generates a population of chromosomes of a predefined size (number of summed contacts columns, as mentioned in step (ii) in "Docking-based comparative molecular contacts analysis (dbCICA)" section in which every chromosome encodes for certain possible column summation model. Chromosomes differ from one another by the set of summed columns and their weights.

Crossover children are the offspring created by selecting vector entries (i.e., genes) from a pair of individual chromosomes in the first generation and combining them to form two complementary children, while mutation children are those created via applying random changes to corresponding parents, i.e., each single parent chromosome is mutated to give a single child by randomly replacing selected gene in the parent chromosome with another from the chromosome population.

Each chromosome is associated with a fitness value that reflects how good the summation of its encoded genes compares to other chromosomes. The fitness functions in dbCICA can be the correlation coefficient (r^2), leave-one-out r^2, or K-fold r^2.

In the current experiment (dbCICA of GK) we implemented a fivefold r^2 as fitness criterion. In this procedure, each chromosome is ranked as follows: The training set is divided into two subsets: fit and test subsets. The test subset is randomly selected to represent *ca.* 20 % of the training compounds. This procedure is repeated over 5 cycles; accordingly, 5 test subsets with their complementary fit subsets are selected for each chromosome (i.e., column summation model). The 5 test subsets should cover *ca.* 100 % of the training compounds by avoiding selecting the same compound in more than one test subset. The fit sets are then utilized to generate 5 sub-models employing the same chromosome. The resulting sub-models are then utilized to predict the bioactivities of the corresponding testing subsets. Finally, the predicted values of all 5 test subsets are correlated with their experimental counterparts to determine the corresponding fivefold r^2.

The following parameters were chosen for GA genetic manipulation in dbCICA of GK: Size of chromosome population = 200; Rate of mating (crossover fraction) = 80 %; Elite count = 1; Maximum number of generations which is needed to exit from GA iteration cycles and completion of the algorithm = 2,000.

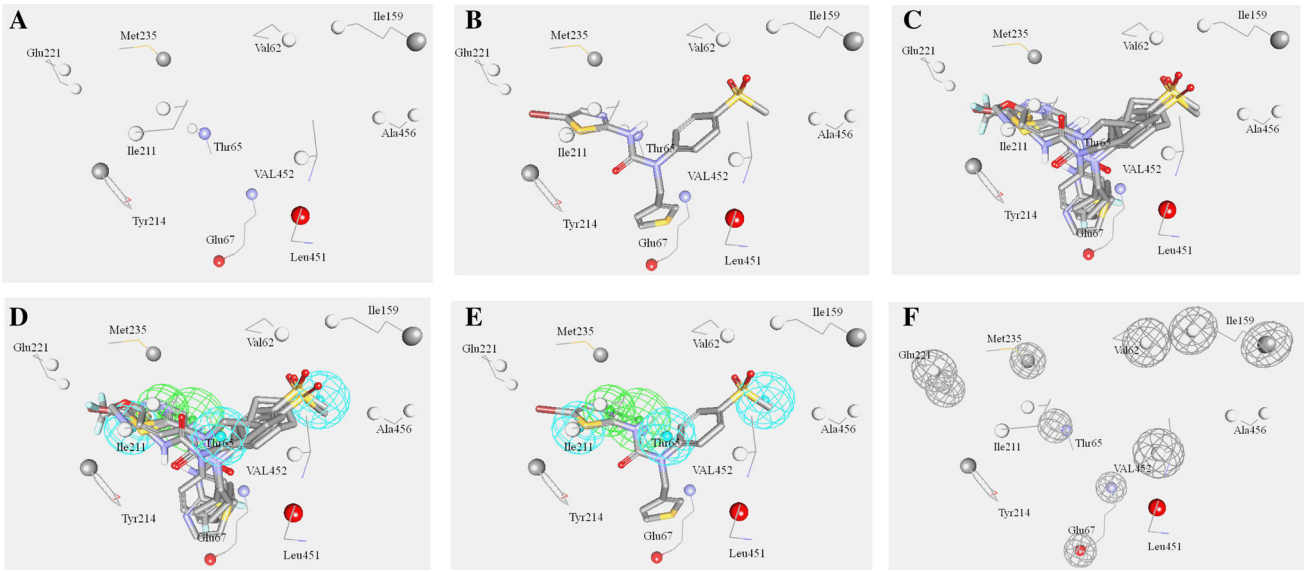


Fig. 3 Steps for manual generation of binding hypothesis HypoA-II as guided by dbCICA model A-II (Tables 3, 4): **a** The binding site moieties in dbCICA model A-II with significant contact atoms shown as *spheres*. **b** The docked pose of the well-behaved compound **26** ($EC_{50} = 6.5$ μM) within the binding pocket, **c** the docked poses of the well-behaved and potent compounds **2, 5, 8 and 26**, **d** manually placed pharmacophoric features onto chemical moieties common among docked well-behaved potent compounds **2, 5, 8 and 26, e** The docked pose of **26** and how it relates to the proposed pharmacophoric features. **f** Exclusion spheres fitted against binding site atoms showing negative correlations with bioactivity (as emergent in dbCICA model A-II)

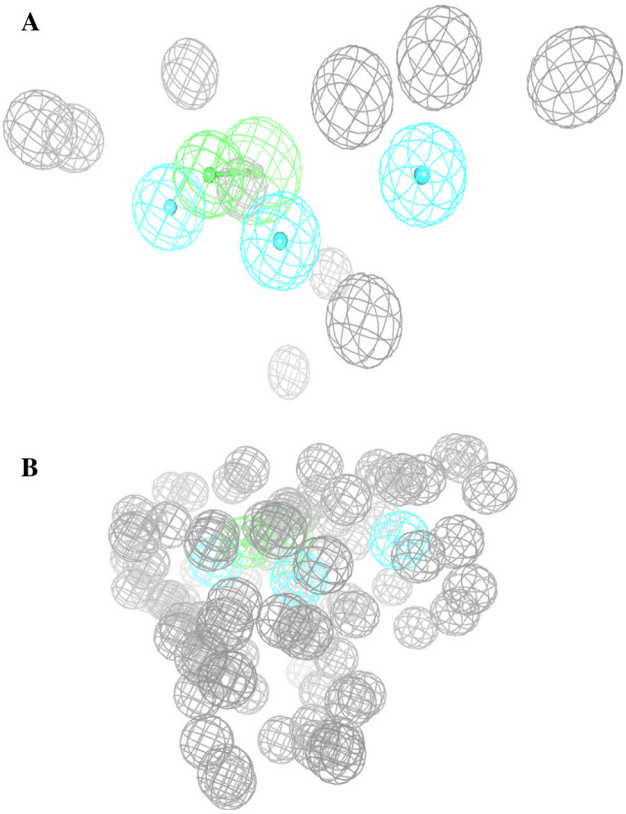


Fig. 4 **a** HypoA-II pharmacophoric features and exclusion spheres *light blue spheres* represent hydrophobic features, *vectored green spheres* represent hydrogen bond acceptor features and *grey spheres* represent exclusion regions, **b** refined HypoA-II with 76 added exclusion spheres

Based on these settings, the numbers of each type of children in the offspring generation is as follows: There is 1 elite child (corresponding to the individual in the parents' generation with the best fitness value), and there are 199 individual children other than the elite child. The algorithm rounds 0.8 (crossover fraction) × 199 = 159.2 to 159 to get the number of crossover children and the remaining 40 (i.e., 199–159) are the mutation population. The elite child is passed to the offspring population without alteration.

Generation of pharmacophores corresponding to successful dbCICA models

In order to utilize dbCICA modeling for effective drug discovery, optimal dbCICA models were used to guide development of pharmacophoric models to be subsequently used as search queries for the discovery of new GKAs. Pharmacophoric models were developed through the following steps:

1. The docking configurations that yielded the best dbCICA models were selected, i.e., LigandFit docking into anhydrous binding site and using unionized ligands via Ligscore1 scoring function (see Tables 2, 3 in "Results and discussion" section). The corresponding docked poses/conformers of the most potent compounds ($EC_{50} < 3$ μM) were retained in the binding pocket while other less potent compounds were discarded.

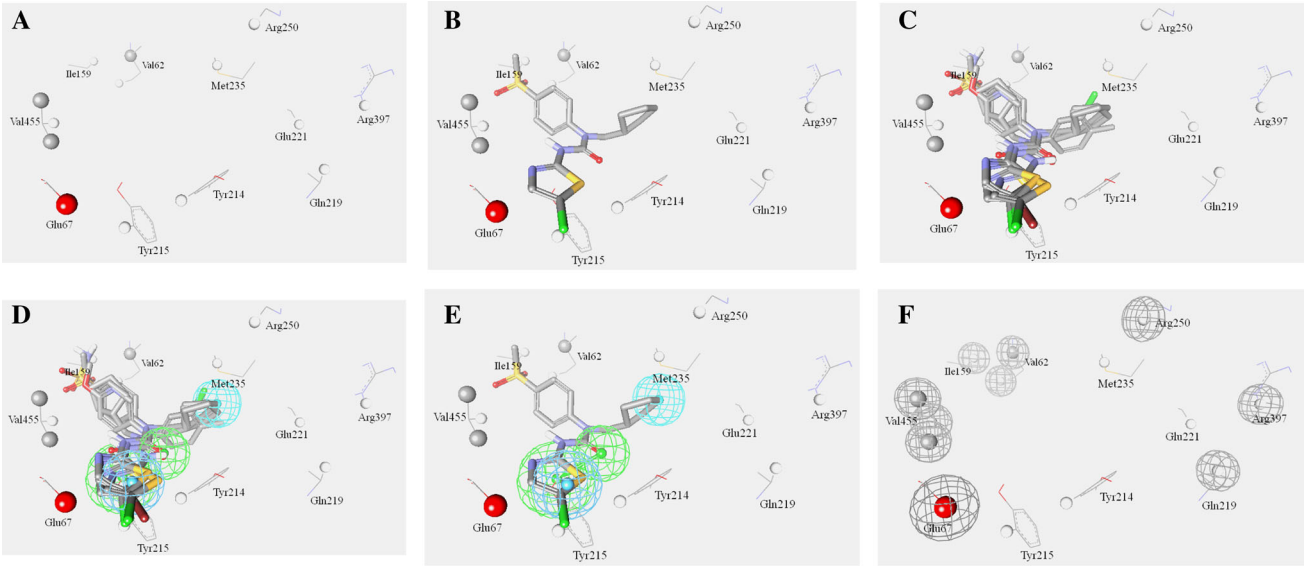


Fig. 5 Steps for manual generation of binding hypothesis HypoA-III as guided by dbCICA model A-III (Tables 3, 4): **a** the binding site moieties in dbCICA model A-III with significant contact atoms shown as *spheres*. **b** The docked pose of the well-behaved compound **3** (EC_{50} = 1.5 μM) within the binding pocket, **c** the docked poses of the well-behaved and potent compounds **3, 10, 13** and **18**. **d** Manually placed pharmacophoric features onto chemical moieties common among docked well-behaved potent compounds **3, 10, 13** and **18**. **e** The docked pose of **3** and how it relates to the proposed pharmacophoric features. **f** Exclusion spheres fitted against binding site atoms showing negative correlations with bioactivity (as emergent in dbCICA model A-III)

2. Subsequently, the best dbCICA model (models 1, 2 or 3, Tables 2, 3) was used to predict the bioactivity of potent compounds in the binding pocket, i.e., by substituting the number of contacts of each docked compound in the regression equation corresponding to the dbCICA model. Well-behaved potent compounds were retained in the binding pocket for subsequent manipulation. Well-behaved compounds are defined as those training compounds that have their bioactivities well-predicted by the selected optimal dbCICA model, i.e., they have the least residual difference between fitted and experimental bioactivities as predicted by the particular dbCICA model.
3. Significant positive contacts in the binding pocket, i.e., those that have weights of 2 or 3, were marked and carefully assessed to identify their closest ligands' moieties. Consensus among potent, well-behaved training compounds to place moieties of common physicochemical properties adjacent to significant contact atom (as defined by the dbCICA model) warrants placing a corresponding pharmacophoric feature onto that region. For example, if potent, well-behaved docked compounds agreed on placing aromatic rings adjacent to certain dbCICA significant contact point (within the predefined distance threshold, see point (i) in "Docking-based comparative molecular contacts analysis (dbCICA)" section then a hydrophobic aromatic feature is placed on top of the aromatic rings. The Pharmacophoric query features were added manually from DS 2.0 feature library and employing default feature radii (1.6 Å). It has to be emphasized that emergence of certain significant contact atom in the binding site is not necessarily indicative of significant ligand interaction(s) with that particular atom in the binding pocket, albeit indicates significant interaction in the vicinity of that atom.
4. Finally, to account for the steric constraints of the binding pocket, binding site atoms that exhibit contacts of negative correlations with bioactivity (i.e., those inverted in step ii, "Docking-based comparative molecular contacts analysis (dbCICA)" section were marked and used as centers for exclusion spheres. Negative contacts identify spaces occupied by docked conformers/poses of inactive compounds and free from active ones and therefore can be filled with exclusion volumes. Exclusion spheres were added manually from DS 2.0 feature library and employing default feature radii (1.2 Å).

Three pharmacophoric hypotheses were chosen to represent the best dbCICA models in the current project.

Molecular field analysis

Comparative molecular field analysis (CoMFA) was performed to assess the ability of corresponding successful

Table 5 Pharmacophoric features, corresponding tolerances and 3D coordinates (X, Y, Z) of optimal dbCICA-based pharmacophore models generated from compounds (**1–41**, Table 1)

Model[a]	Definitions	Chemical features				
		HBA[c]		Hbic[d]	HbicArom[e]	Hbic
HypoA-If	Tolerances[b]	1.6 (tail)	2.2 (head)	1.6	1.6	1.6
	Coordinates					
	X	34.49	32.07	41.17	38.38	40.13
	Y	20.17	19.59	11.57	17.58	15.87
	Z	59.15	57.48	63.23	61.57	61.51

Model[a]	Definitions	Chemical features				
		Hbic	Hbic	Hbic	HBA	
HypoA-IIg	Tolerances	1.6	1.6	1.6	1.6 (tail)	2.2 (head)
	Coordinates					
	X	43.50	36.04	39.76	42.30	43.18
	Y	12.08	19.50	16.19	12.98	15.07
	Z	62.32	59.63	62.67	60.82	58.86

Model[a]	Definitions	Chemical features			
		Hbic	HbicArom	HBA	
HypoA-IIIh	Tolerances	1.6	1.6	1.6 (tail)	2.2 (head)
	Coordinates				
	X	42.55	45.51	42.89	41.68
	Y	11.21	17.99	15.09	17.26
	Z	61.67	64.90	64.53	66.20

[a] Pharmacophoric hypothesis shown in Figs. 2, 4 and 6

[b] Tolerances: refer to the radius of feature spheres (Å)

[c] *HBA* Hydrogen bond acceptor feature

[d] *Hbic* Hydrophobic feature

[e] *HbicArom* Hydrophobic aromatic feature

[f] Number of exclusion spheres in HypoA-I before HipHop-steric refinement = 10 of 12 Å tolerance, at the following X,Y,Z coordinates: (35.42, 14.70, 58.90), (36.36, 13.42, 61.95), (40.59, 21.82, 3.78), (33.32, 16.79, 63.23), (34.51, 15.24, 64.74), (40.98, 14.39,57.13), (41.17, 8.18, 60.03), (31.72,17.17, 60.14), (42.35, 5.79, 63.17), and (44.66, 10.45, 56.58)

[g] Number of exclusion spheres in HypoA-II before HipHop-steric refinement = 10 of 12 Å tolerance, at the following X,Y,Z coordinates: (42.11, 6.85, 63.12), (44.15, 7.40, 61.89), (44.67, 20.99, 60.14),(47.71, 21.44, 64.30), (30.99, 20.52, 58.41), (32.55, 16.82, 59.53), (40.86, 10.55, 58.62), (42.85, 14.58, 60.16), (33.94, 15.14, 61.15), and (36.30, 18.49, 64.90)

[h] Number of exclusion spheres in HypoA-III before HipHop-steric refinement = 10 of 12 Å tolerance, at the following X,Y,Z coordinates: (64.95, −77.00, 79.71), (62.54, −81.31, 79.06), (60.55, −75.64, 80.26), (73.04, −71.68, 82.70), (59.35, −72.78, 67.97), (56.98, −72.32, 69.30), (58.82, −71.09, 67.90), (68.13, −77.24, 76.84), (63.82, −66.87, 67.10), (65.16, −75.88, 69.98)

dbCICA-based molecular alignments to give self-consistent CoMFA models [64, 82, 83]. We used Molecular Field Analysis (MFA) and G/PLS modules implemented in CERIUS2® to perform 3D QSAR analysis [94]. The alignments of different activators came directly from the top-scoring conformers/orientations according to dbCICA-based highest ranking docking/scoring combinations (e.g., Tables 3, 4). For each alignment, the interaction fields between the ligands and proton (positively charged), hydrogen-bond donor/acceptor and methyl (neutral) probes were calculated employing a regularly spaced rectangular grid of 2.0 Å spacing. The spatial limits of the molecular field were defined automatically, and were extended past the van der Waals volume of all the molecules in the X, Y and Z directions. The ligands were assigned partial charges using the Gasteiger method implemented within CERIUS2®. The energy fields were calculated employing the default UNIVERSAL force field® (version 1.02) implemented within CERIUS2® [95], and were truncated to ±50 kcal/mol. For compounds (**1–41**, Table 1) the calculation gave range from 2,376 to 4,320 variables for each compound (792–1,440 variable/probe). For the second set

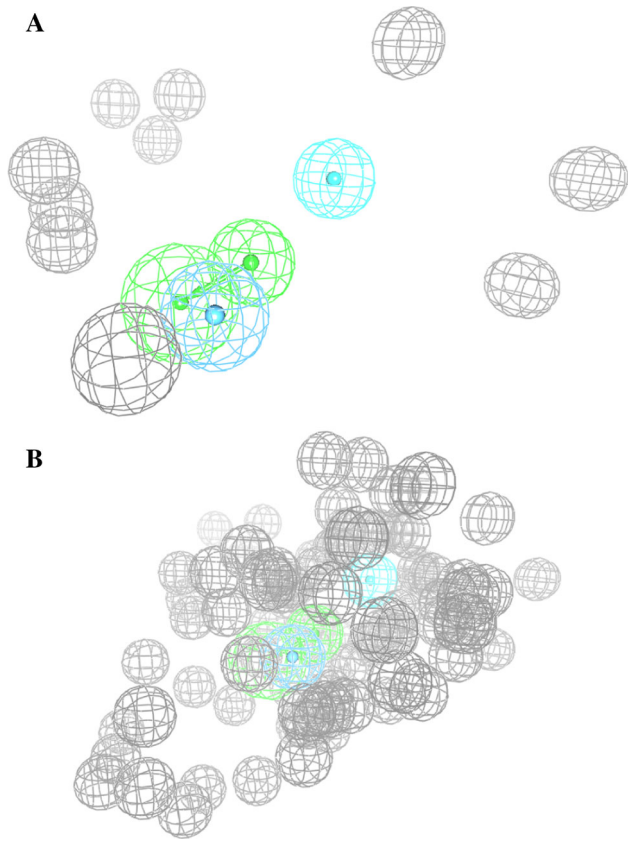


Fig. 6 **a** HypoA-III pharmacophoric features and exclusion spheres *light blue spheres* represent hydrophobic features, *dark blue spheres* represent hydrophobic aromatic features, *vectored green spheres* represent hydrogen bond acceptor features and *gray spheres* represent exclusion regions, **b** refined HypoA-III with 80 added exclusion spheres

of activators (**42–71**, Table 1), the calculation gave range from 2,640 to 4,320 variables for each compound (with 880–1,440 variable/probe), while.

To derive the best possible 3D QSAR statistical model for each docking/scoring combination, we used Genetic Partial Least Squares (G/PLS) analysis to search for optimal regression equations capable of correlating the variations in biological activities of the training compounds with variations in the corresponding interaction fields [94]. G/PLS is derived from two methods: genetic function approximation (GFA) and partial least squares (PLS). GFA techniques rely on the evolutionary operations of "crossover and mutation" to select optimal combinations of descriptors (i.e., chromosomes) capable of explaining bioactivity variation among training compounds from a large pool of possible descriptor combinations (i.e., chromosomes population). Each chromosome is associated with a fitness value that reflects how good it is compared to other solutions. The fitness function employed herein is based on Friedman's 'lack-of-fit' (LOF) [94].

Table 6 Training subset used for adding excluded spheres for dbCICA-based hypotheses using Hiphop-Refine module of CATALYST for HypoA-I, HypoA-II, HypoA-III, HypoB-I, HypoB-II and HypoB-III

Compound[a]	EC50 (μM)	Principal[b]	MaxOmitFeat[c]
42	3.20	0	2
43	>10	0	1
44	>10	0	2
45	>10	0	1
46	>10	0	1
47	>10	0	1
48	4.22	0	1
49	>10	0	1
50	0.91	1	0
51	>10	0	0
52	0.57	1	0
53	0.40	1	0
54	0.09	2	0
55	1.33	1	0
56	0.65	1	0
57	0.17	2	0
58	1.26	1	0
59	1.78	1	0
60	0.13	2	0
61	0.09	2	0
62	0.29	1	0
63	0.42	1	0
64	0.10	2	0
65	0.11	2	0
66	0.95	1	0
67	0.61	1	1
68	5.51	0	1
69	0.02	2	0
70	0.09	2	0
71	0.03	2	0

[a] Compound number as in Table 1

[b] Principal value according to activity (see Receiver operating characteristic (ROC) curve analysis section under Experimental)

[c] Maximum omitted feature (see Receiver operating characteristic (ROC) curve analysis section under Experimental)

G/PLS algorithm uses GFA to select appropriate basis functions to be used in a model of the data, and PLS regression as the fitting technique to weigh the basis functions' relative contributions in the final model. Application of G/PLS allows the construction of larger QSAR equations while avoiding overfitting and eliminating most variables [94].

Our preliminary diagnostic trials suggested the following optimal G/PLS parameters: Explore linear equations of 4–8 terms at mating and mutation probabilities

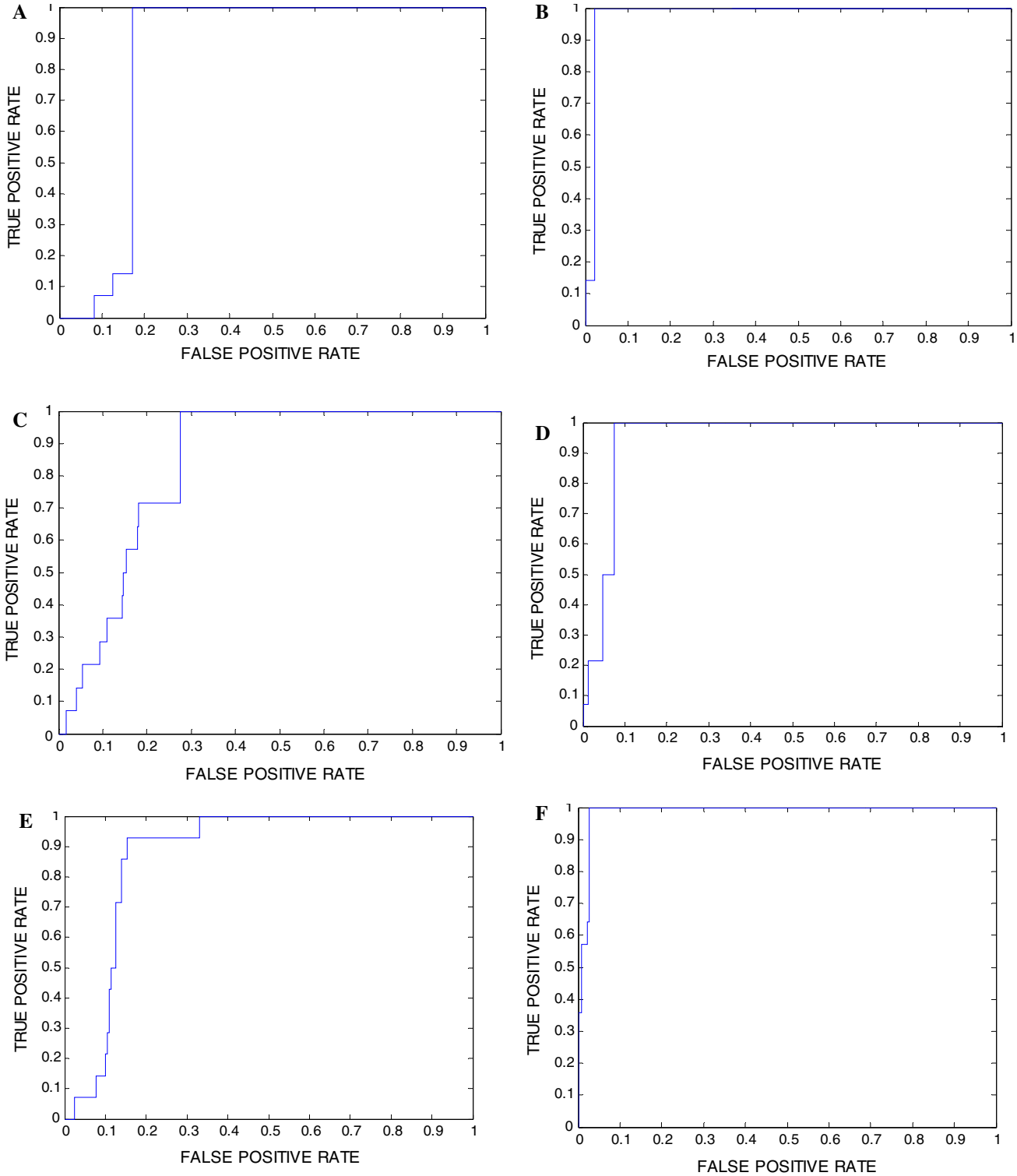


Fig. 7 Receiver operating characteristic (ROC) curves of dbCICA-based **a** HypoA-I, **b** sterically refined HypoA-I, **c** HypoA-II, **d** sterically refined HypoA-II, **e** HypoA-III and **f** sterically refined HypoA-III

of 50 %; population size = 500; number of generations (iterations) = 30,000 and LOF smoothness parameter = 1.0. However, the optimal number of PLS latent variables (or principle components) was determined for each CoMFA model through assessing the corresponding predictive r^2 (r^2_{PRESS}) calculated from a test set of 8

Table 7 Performance of dbCICA-selected pharmacophores as 3D search queries

Pharmacophore model	ROC[a]–AUC[b]	ACC[c]	SPC[d]	TPR[e]	FNR[f]
HypoA-I	84	96	99	14	1
HypoA-II	84	96	97	71	3
HypoA-III	87	96	96	100	4
Refined-HypoA-I	98	96	99	14	1
Refined-HypoA-II	95	96	98	50	2
Refined-HypoA-III	99	96	98	64	2

[a] *ROC* receiver operating characteristic

[b] *AUC* area under the curve (%)

[c] *ACC* overall accuracy (%)

[d] *SPC* overall specifity (%)

[e] *TPR* Overall true positive rate

[f] *FNR* Overall false negative rate (%)

Table 8 The statistical results of the CoMFA models obtained via dbCICA-based docking/scoring combinations (in Table 3)

dbCICA models[a]	Docking conditions[a]	Scoring function[a]	Terms[b]	PC[c]	Statistical criteria		
					r^{2}[d]	r^2_{BS}[e]	r^2_{PRESS}[f]
Highest[g] ranking							
A-I[g]	Unionized Ligands-anhydrous binding pocket	Ligscore1	8	4	0.915	0.857	0.729[i]
A-II	Unionized Ligands-anhydrous binding pocket	PLP2	8	3	0.906	0.402	0.515[j]
A-III	Unionized Ligands-anhydrous binding pocket	PMF	6	5	0.917	0.863	0.569[k]
Lowest ranking[h]	Unionized ligands-hydrous binding pocket	PLP1	4	3	0.652	0.548	0.156[l]

[a] dbCICA models and corresponding docking-scoring conditions were selected from Tables 2 and 3

[b] Number of CoMFA descriptors in the best 3D-QSAR model

[c] Number of principal components (latent variables) in the best 3D-QSAR model

[d] Non-crossvalidated correlation coefficient for 33 training compounds

[e] Bootstrapping correlation coefficient

[f] Predictive r^2 determined for the 8 test compounds

[g] Docking-scoring conditions of highest ranking dbCICA models. Models numbered as in Table 3

[h] Docking-scoring conditions of lowest ranking dbCICA models selected from Table 2

[i] r^2_{PRESS} value after removing one outlier (**17**, Table 1)

[j] r^2_{PRESS} value after removing one outlier (**4**, Table 1)

[k] r^2_{PRESS} value after removing one outlier (**17**, Table 1)

[l] r^2_{PRESS} value after removing one outlier (**17**, Table 1)

GKAs in the first unionized activators and a set of six activators for the second ionized set (labeled in Table 1). Test molecules were selected by ranking compounds **1–41** and **42–71** according to their EC_{50} values followed by selecting every fifth compound for the test set starting from the high potency end. Test molecules were aligned according to the particular docking/scoring configuration, and their activities were predicted by corresponding G/PLS models generated from the training set (33 compounds for case one and 24 compounds for the second case) and employing a range of 3–6 latent variables. The optimum number of principle components was defined as the one leading to the highest predictive r^2_{PRESS} and lowest sum of squared deviations between predicted and actual activity values for every molecule in the test set (PRESS). Predictive r^2_{PRESS} is defined as [64, 82, 83]:

$$r^2_{PRESS} = (SD - PRESS)/SD \quad (1)$$

Where SD is the sum of the squared deviations between the biological activities of the test set and the mean activity of the training set molecules.

Table 9 In silico hits with their fit values against (HypoA-I) and their in vitro GK activation folds

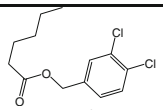
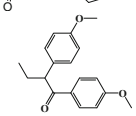
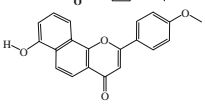
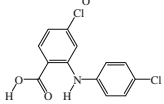
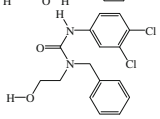
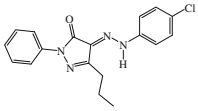
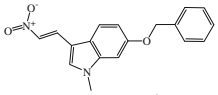
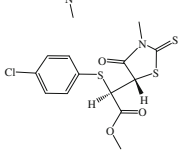
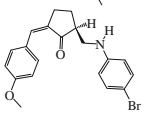
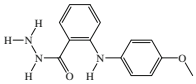
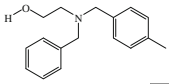
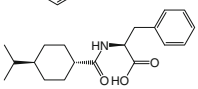
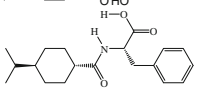
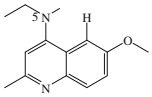
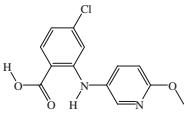
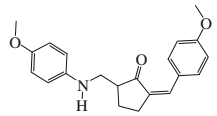
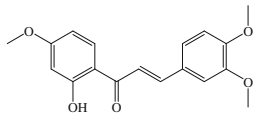
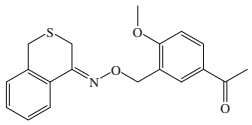
No.	NCI code or Chemical Name	Structure	dbCICA contacts[a]	Experimental Fold Activation at 10 μM
72	7032		12	4
73	63378		14	1
74	123386		16	5
75	158696		15	6.7[b]
76	204740		15	4
77	270451		13	1
78	294438		14	0
79	303571		13	1
80	661235		13	5
81	151943		12	5
82	205363		11	1
83	Nateglinide		12	1.1
84	Secbumeton		11	0.5
85	16122		15	6[b]
86	60303		15	1

Table 9 continued

No.	NCI code or Chemical Name	Structure	dbCICA contacts[a]	Experimental Fold Activation at 10 μM
87	661234		11	4
88	91849		13	0
89	319449		11	1

[a] Contacts summations according to dbCICA model **I** (Tables 2, 3)

[b] Activation folds values were measured as duplicated readings

All 3D QSAR models were cross-validated employing leave-one-out (LOO) cross-validation and bootstrapping [64, 82, 83, 94].

Receiver operating characteristic (ROC) curve analysis

For both cases, the manually-prepared pharmacophoric hypotheses (i.e., LigandFit-based) were validated by assessing their abilities to selectively capture diverse GK active compounds from a large test list of actives and decoys. The testing decoy list was prepared as described by Verdonk et al. [15, 97]. For each active compound in the test set, around 33 decoys were randomly chosen from the ZINC database [96]. See section SM-3 in the supplementary material for detailed experimental and theoretical explanations of decoy list generation and ROC analysis.

Addition of exclusion volumes

dbCICA-based pharmacophore models HypoA-I, HypoA-II and HypoA-III obtained from the first set of compounds (**1–41**, Table 1), (Figs. 1, 3, 5) were decorated with exclusion volumes employing HipHop-Refine module of Catalyst. The same was applied for the pharmacophoric models HypoB-I, HypoB-II and HypoB-III obtained for the second set (**42–71**, Table 1), (Figs. 1, 3, 5). HipHop-Refine uses inactive training compounds to construct excluded volumes that resemble the steric constrains of the binding pocket. It identifies spaces occupied by the conformations of inactive compounds and free from active ones. These regions are then filled with excluded volumes [98–102]. Compounds **42–71** (Table 1) were used for constructing appropriate exclusion regions around all dbCICA-based pharmacophores (HypoA-I, HypoA-II, HypoA-III, HypoB-I, HypoB-II and HypoB-III). Table 6 shows the training subset used for HipHop-Refine modeling.

In HipHop-Refine the user defines how many molecules must map the particular hypothesis completely or partially through the principal and maximum omitted features (MaxOmitFeat) parameters. Active compounds are normally assigned MaxOmitFeat parameter of zero and principal value of 2 to instruct the software to consider all their chemical moieties in HipHop-Refine modeling and to fit them against all pharmacophoric features of a particular hypothesis. On the other hand, inactive compounds are allowed to miss one (or more) features by assigning them a MaxOmitFeat of 1 (or 2) and principal value of zero [98–102].

We decided to consider 0.40 μM as an appropriate activity/inactivity cutoff threshold. Accordingly, activators of EC_{50} values <0.40 μM were regarded as "actives" and were assigned principal and MaxOmitFeat values of 2 and 0, respectively. On the other hand, activators of EC_{50} values ranging from 0.40 to 3.00 μM were considered as intermediates and were assigned Principal values of 1 and MaxOmitFeat parameter of 0 (rarely 1). While activators of EC_{50} values >3.2 μM were regarded as inactives and were assigned a Principal value of 0. However, each inactive compound was carefully evaluated to assess whether its low potency is attributable to missing one or more pharmacophoric features, i.e., compared to active compounds, or related to possible steric clashes within the binding pocket, or due to both factors. Therefore, inactive compounds suspected of missing one or more pharmacophoric features were assigned MaxOmitFeat values of 1 or 2, respectively. Spaces occupied by conformers and/or

Table 10 In silico hits with their fit values against (HypoA-II) and their in vitro GK activation folds

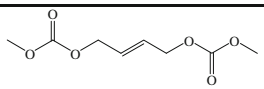
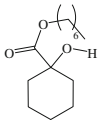
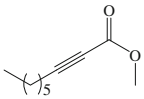
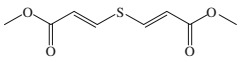
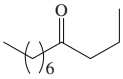
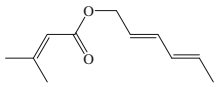
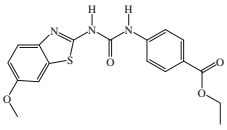
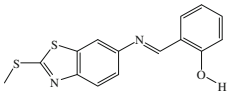
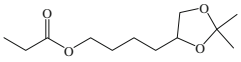
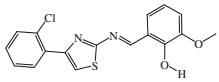
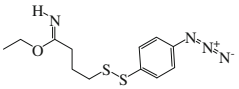
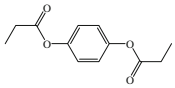
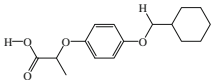
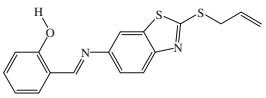
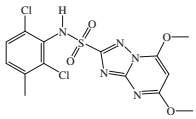
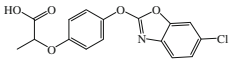
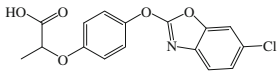
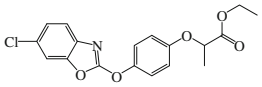
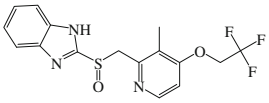
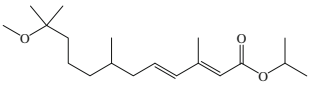
No.	NCI code or Chemical Name	Structure	dbCICA contacts[a]	Experimental Fold Activation at 10 μM
90	561		11	7.4[b]
91	6662		13	6.5[b]
92	72102		11	4
93	122680		11	7.3[b]
94	158509		14	5.5[b]
95	195150		11	4
96	290499		11	6.8[b]
97	327346		11	5
98	40521		10	3
99	338510		11	1
100	370342		12	1
101	55130		11	6.9[b]
103	656091		11	1
104	327349		11	1
105	Metosulam		10	1.3
106	Fenoxaprop		11	1.2

Table 10 continued

No.	NCI code or Chemical Name	Structure	dbCICA contacts[a]	Experimental Fold Activation at 10 μM
107	Fenoxaprop p		10	0.4
108	Fenoxaprop-P-ethyl		10	0.8
109	Lansoprazole		12	1
110	Methoprene		10	0.5

[a] Contacts summations according to dbCICA model II (Tables 2, 3)

[b] Activation folds values were measured as duplicated readings

mappings of this group of compounds and free from conformers and/or mappings of active compounds are filled with excluded volumes.

However, compounds that seem to be inactive mainly due to steric clashes within the binding pocket were assigned MaxOmitFeat value of zero. This value instructs HipHop-Refine to force inactive compound(s) to map all the pharmacophoric features of the binding model, and therefore permits the software to identify spaces occupied by excess structural fragments/features of such inactive compounds and fill them with excluded volumes [100].

HipHop-Refine was configured to allow a maximum of 100 exclusion spheres to be added to the dbCICA pharmacophoric hypotheses. Table 6 (see "Results and discussion" section) shows the training compounds in this step and their corresponding principal and MaxOmitFeat parameters.

In-silico screening of the NCI database for new GK activators

The refined versions of HypoA-I, HypoA-II, HypoA-III, HypoB-I and HypoB-III (Figs. 2b, 4b, 6b, 9b, 13b) were employed as 3D search query to screen the National Cancer Institute (NCI) list of compounds (238,819 compounds) and our *in-house* list of drugs and agrochemicals (3,002 compounds). HypoB-II was excluded due to its inferior ROC behavior compared to HypoB-I and HypoB-III. The screening was performed employing the "Best Flexible Database Search" option implemented within CATALYST module of DS 2.0.

NCI hits were subsequently filtered based on molecular weight, such that only hits of molecular weights ≤500 Da were retained. Surviving hits were docked into GK protein (PDB code: 1V4S) employing the same docking conditions of corresponding dbCICA models. The resulting docked poses were subsequently analyzed for critical contacts according to corresponding successful dbCICA models and the sums of critical contacts for each hit compound were used to rank the corresponding hits and prioritize subsequent in vitro testing. Tables 18 and 19 show the highest ranking hits and their experimental in vitro bioactivities.

In vitro testing of GK activation

Chemicals

All chemicals needed for bioassay were purchased from Sigma-Aldrich Company and were used without further purification.

In vitro assay

Bioassay is based on the phosphorylation of D-glucose by GK to yield D-glucose-6-phosphate, which is oxidized by the enzyme glucose-6-phosphate dehydrogenase (G6PD) in the presence of NADP into 6-phospho-D-gluconate and NADPH. The later has λ max of 340 nm. The rate at which NADPH is generated is directly related to the catalytic activity of GK.

The bioassay procedure was performed as reported previously [103]. Briefly, stock solutions of test samples were prepared in DMSO, and then serially diluted with deionised water to give the desired working concentrations.

Table 11 In silico hits with their fit values against (HypoA-III) and their in vitro GK activation folds

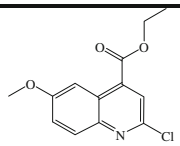
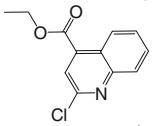
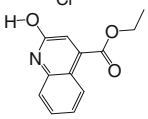
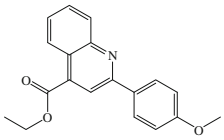
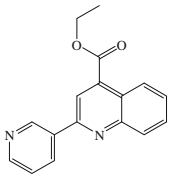
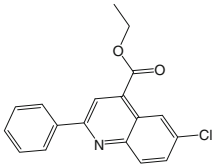
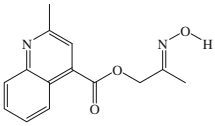
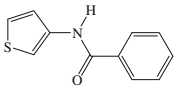
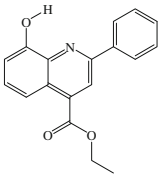
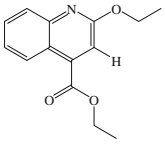
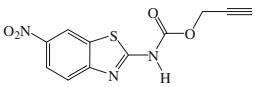
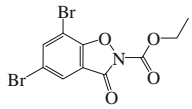
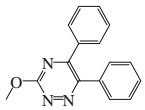
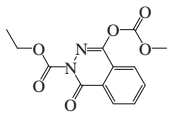
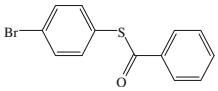
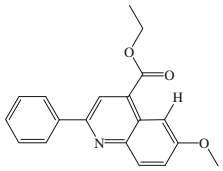
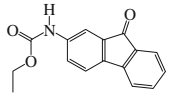
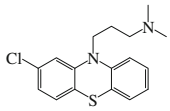
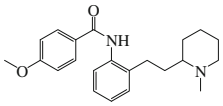
No.	NCI code or Chemical Name	Structure	dbCICA contacts[a]	Experimental Fold Activation at 10 μM
111	1708		12	6.1[b]
112	25659		12	4.5[b]
113	25661		12	5
114	25680		13	4
115	26085		13	6.2[b]
116	42127		12	5.4[b]
117	50102		12	7.1[b]
118	108954		13	7.5[b]
119	117553		13	5
120	26073		12	4
121	327387		10	5

Table 11 continued

No.	NCI code or Chemical Name	Structure	dbCICA contacts[a]	Experimental Fold Activation at 10 μM
122	355381		10	1
123	402878		12	0
124	609357		12	1
125	99115		13	6.5[b]
126	101043		13	5
127	81317		11	5
128	Chlorpromazine		10	1.4
129	Encainide		14	0.9

[a] Contacts summations according to dbCICA model III (Tables 2, 3)

[b] Activation folds values were measured as duplicated readings

Bioassay was performed by adding 3 μL of tested sample solution to a reaction mixture (90 μL) composed of Tris HCl buffer (75 mM, 24 mL, pH 9.0 at 30 °C); $MgCl_2$ (600 mM in deionized water, 1 mL); ATP (120 mM in deionized water, 1 mL); ß-D (+) glucose (360 mM in deionized water, 1 mL) and NADP (27 mM in deionized water, 1 mL). Subsequently, G6PD (1,000 units/mL in cold deionized water, 3 μL) was added followed by human GK solution (0.05 units/mL, 3 μL) in cold tris buffer (pH 8.5, 4 °C) to initiate the reaction. The samples' concentrations were fixed at 10 μM in the reaction well. The change in absorbance at λ 340 nm is measured. The rate of enzyme reaction was considered as the reference for activation process. Change in absorbance (Rate) was determined at 5, 10 and 15 min for all tested compounds.

Activation of human GK was calculated as percent activity of the unactivated enzyme control. DMSO concentrations were kept <1 % in all experiments and controls. Some samples were prepared in duplicates.

Results and discussion

Basic concept of dbCICA

Although docking engines suffer from being unable to calculate free energy of binding, they normally succeed in reproducing co-crystallized ligand poses/conformations among their high-ranking docking solutions [17, 21, 23, 47, 61, 90]. This suggests that its quite possible to correlate docked 3D ligand poses and conformers with bioactivities, as shown previously using CoMFA modeling [64, 82, 83].

In dbCICA, the interest is focused on identifying a set of atoms within the binding site that tend to contact with

Table 12 Different liganfit-based docking conditions, their corresponding best dbCICA parameters and statistical criteria for compounds (**42–71**, Table 1)

Docking conditions			Optimal dbCICA parameters			dbCICA statistical criteria		
Ligands' ionization state	Explicit water[a]	Scoring functions	Contacts distance threshold (Å)[b]	Number of positive contacts[c]	Number of negative contacts[d]	R_{30}[e]	r^2_{LOO}[f]	r^2_{5-fold}[g]
Ionized	Present	Jain	3.5	5	5	0.83	0.81	0.81
		Ligscore 1	2.5	9	10	0.79	0.76	0.74
		Ligscore 2	3.5	9	10	0.76	0.73	0.73
		PLP1	2.5	10	10	0.86	0.84	0.84
		PLP2	3.5	5	5	0.85	0.83	0.83
		PMF	3.5	9	5	0.77	0.73	0.76
	Absent	*Jain*[h]	*3.5*	*10*	*10*	*0.89*	*0.88*	*0.89*
		Ligscore 1	2.5	4	10	0.8	0.77	0.79
		Ligscore 2	2.5	8	5	0.78	0.76	0.76
		PLP1	3.5	4	2	0.86	0.84	0.84
		PLP2	3.5	5	10	0.84	0.82	0.83
		PMF[h]	*3.5*	*9*	*10*	*0.91*	*0.9*	*0.9*
Unionized	Present	Jain	3.5	5	5	0.85	0.83	0.84
		Ligscore 1	2.5	6	10	0.81	0.78	0.77
		Ligscore 2	2.5	6	10	0.85	0.83	0.84
		PLP1	3.5	10	10	0.84	0.81	0.83
		PLP2[h]	*3.5*	*8*	*5*	*0.89*	*0.88*	*0.87*
		PMF	2.5	10	10	0.82	0.8	0.8
	Absent	Jain	3.5	6	5	0.76	0.73	0.75
		Ligscore 1	2.5	5	10	0.85	0.82	0.83
		Ligscore 2	2.5	7	5	0.79	0.76	0.77
		PLP1	2.5	8	10	0.86	0.84	0.85
		PLP2	2.5	6	10	0.86	0.84	0.84
		PMF	3.5	10	10	0.81	0.78	0.79

[a] Crystalographically explicit water of hydration

[b] Distance thresholds used to define ligand-binding site contacts

[c] Optimal number of combined (i.e., summed) bioactivity-enhancing ligand/binding site contacts

[d] Optimal number of combined (i.e., summed) bioactivity-disfavoring ligand/binding site contacts

[e] Non-cross-validated correlation coefficient for 30 training compounds

[f] Cross-validation correlation coefficients determined by the leave-one-out technique

[g] Cross-validation correlation coefficients determined by the leave-20 %-out technique repeated 5 times

[h] Italic parameters correspond to the best docking/scoring combinations in LigandFit-based docking

Table 13 Highest ranking dbCICA models, their corresponding parameters and statistical criteria

dbCICA model	Ligands' ionization state	Explicit water[a]	Scoring function	Contacts distance threshold[b]	Number of positive contacts[c]	Number of negative contacts[d]	R_{24}[e]	r^2_{LOO}[f]	r^2_{5-fold}[g]	F statistic
B-I	Ionized	Absent	JAIN	3.5	10	10	0.89	0.88	0.89	221.73
B-II	Ionized	Absent	PMF	3.5	9	10	0.91	0.90	0.90	274.14
B-III	Unionized	Present	PLP2	3.5	8	5	0.89	0.88	0.87	234.87

The presented information in this table are extracted from Table 12 except for F-statistic, which is calculated from the correlation connecting $-\log(EC_{50})$ and the contacts sums of the corresponding dbCICA models

[a] Crystalographically explicit water of hydration

[b] Distance thresholds used to define ligand-binding site contacts

[c] Optimal number of combined (i.e., summed) bioactivity-enhancing ligand/binding site contacts

[d] Optimal number of combined (i.e., summed) bioactivity-disfavoring ligand/binding site contacts

[e] Non-cross-validated correlation coefficient for 30 training compounds

[f] Cross-validation correlation coefficients determined by the leave-one-out technique

[g] Cross-validation correlation coefficients determined by the leave-20 %-out technique repeated 5 times

Table 14 Critical binding site contact atoms proposed by optimal dbCICA models generated from compounds (**42–71**, Table 1)

dbCICA model[a]	Favored contact atoms (positive contacts)[b]		Disfavored contact atoms (negative contacts)[e]
	Amino acids (or water) and corresponding atom identities[c]	Weights[d]	
B-I	ARG63:O	1	GLU221:OE1; GLU96:O; ILE211:HD13; LYS459:CG; LYS459:HB1; MET210:HG2; THR65:N; TYR215:CE1; VAL455:HG22; VAL62:CG1
	GLU67:HA	2	
	ILE159:HG21	3	
	MET210:CE	3	
	MET210:O	2	
	MET235:HE3	1	
	MET235:SD	2	
	TYR215:CD1	1	
	TYR61:C	1	
	VAL62:HA	3	
B-II	ARG63:N	2	ALA456:CB; ARG250:NH1; LEU451:O; SER69:HB1; THR65:HG21; THR65:N; VAL452:CA; VAL452:CG2; VAL455:HG23; VAL62:CG1
	ILE211:N	3	
	LEU451:HB2	3	
	MET210:CE	3	
	MET235:HE2	3	
	THR65:C	1	
	THR65:HA	2	
	TYR215:HE2	1	
	GLY97:C	2	
B-III	ARG63:C	3	ALA201:HB3; ARG250:NH1; ARG63:HD2; :VAL452:C; VAL452:CG1
	GLN98:HE22	1	
	GLY68:O	2	
	LEU451:HB1	3	
	THR65:HG23	1	
	TYR214:HD1	2	
	PRO66:CG	3	
	ASP158:OD2	1	

[a] As in Table 13

[b] Bioactivity-proportional ligand/binding site contacts

[c] Binding site amino acids and their significant atomic contacts Atom codes are as provided by the protein data bank file format (e.g., ARG63:O encodes for oxygen atom (O) of arginine number 63) except for hydrogen atoms which were coded by DS 2.0

[d] Degree of significance (weight) of corresponding contact atom

[e] Bioactivity-disfavoring ligand/binding site contacts

potent docked ligands while avoid poorly active docked ligands. If such a set of contact atoms is identified for a docked list of ligands, then one can assume that the docking configuration is successful, i.e., it arranged the molecules in a such way that can explains variation in bioactivity.

High ligand-receptor affinity is mediated by certain critical number of interactions. However, the fact that docking engines and scoring functions evaluate large number of ligand-receptor interactions to generate their docking solutions means the influence of critical interactions might be unnoticed. In this context, identifying a set affinity-discriminating contact atoms within the binding site should help not only to validate a particular docking configuration, but also to highlight the critical ligand-receptors interactions responsible for affinity as such discriminatory contacts encode for nearby attractive interactions. In fact, the resulting dbCICA models can be translated into abstract pharmacophoric models of limited number of critical binding features that can be used as 3D search queries to mine for new ligands.

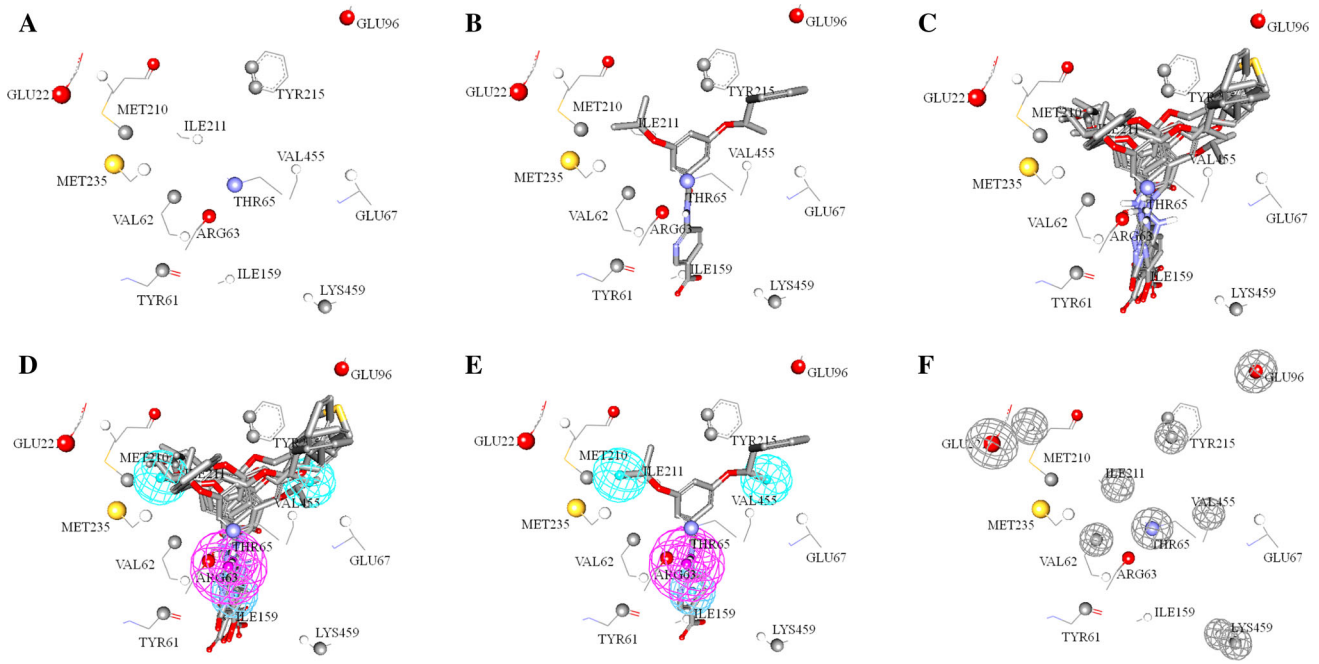


Fig. 8 Steps for manual generation of binding hypothesis HypoB-I as guided by dbCICA model B-I (Tables 13, 14): **a** the binding site moieties in dbCICA model B-I with significant contact atoms shown as *spheres*. **b** The docked pose of the well-behaved compound **67** ($EC_{50} = 0.02$ μM) within the binding pocket, **c** the docked poses of the well-behaved and potent compounds **60, 61, 62, 69, 70** and **71**. **d** Manually placed pharmacophoric features onto chemical moieties common among docked well-behaved potent compounds **60, 61, 62, 69, 70** and **71**. **e** The docked pose of **69** and how it relates to the proposed pharmacophoric features. **f** Exclusion spheres fitted against binding site atoms showing negative correlations with bioactivity

Practical aspects and implementation of dbCICA

dbCICA concept requires a clear definition of 'contacts', i.e., the inter-atomic distance thresholds that can be considered as reasonable contacts. The fact that most ligand-receptor interactions, e.g., hydrogen-bonding and van der Waals' forces, illustrate optimal strength at distance range of 2.5–3.5 Å, prompted us to perform dbCICA analysis by implementing two distance thresholds as inter-atomic contacts, i.e., 2.5 or 3.5 Å. Inter-atomic distances $\leq$ the predefined threshold are considered as contacts and are encoded by a binary value of one, while larger distances are considered as non-contacts and encoded by zeros.

As mentioned earlier, discriminatory ligand-receptor contacts are surrogates of nearby critical ligand-receptor interactions. However, since ligand-receptor affinity is normally mediated by a set of concomitant critical attractive interactions, it is expected that their proxy contacts are also concomitant. This basic principle is represented in dbCICA analysis by searching for discriminatory ligand-receptor contacts that have their *summation* values directly proportional to bioactivity, i.e., search of concurrent contacts rather than separate contacts.

Furthermore, our dbCICA algorithm has the option of allowing any particular positive contacts column to emerge up to three times in the optimal summation model, i.e., allows variable weights for contacts. This option identifies intermolecular contacts of higher weights or contributions in the optimal dbCICA models (see point (iii) under "Docking-based comparative molecular contacts analysis (dbCICA)" section.

Still, some contacts are expected to be inversely proportional to bioactivity (i.e., negative contacts). These contacts encode for repulsive interactions (steric clashes). However, to remove any correlation faults resulting from summations of negative and positive contacts during search for optimal contact combinations, dbCICA analysis is preceded by scanning the correlations between each contact column and bioactivity. Inversely proportional contacts, i.e., negative contacts, are removed during the initial search phase for optimal combinations of positive contacts.

Nevertheless, since ligand-receptor binding is controlled by both attractive and repulsive forces, both positive and negative contacts are later combined in dbCICA models. However, to maintain the consistency of correlation calculations, i.e., maintain the trend of direct proportionality between contacts combinations and bioactivities, it was decided to invert negative contacts, i.e., by converting their zeros to ones and vice versa. Subsequently, a second search phase is performed to find optimal summations of negative contacts that upon combination with previously defined optimal positive contacts summations yield optimal correlations with bioactivity.

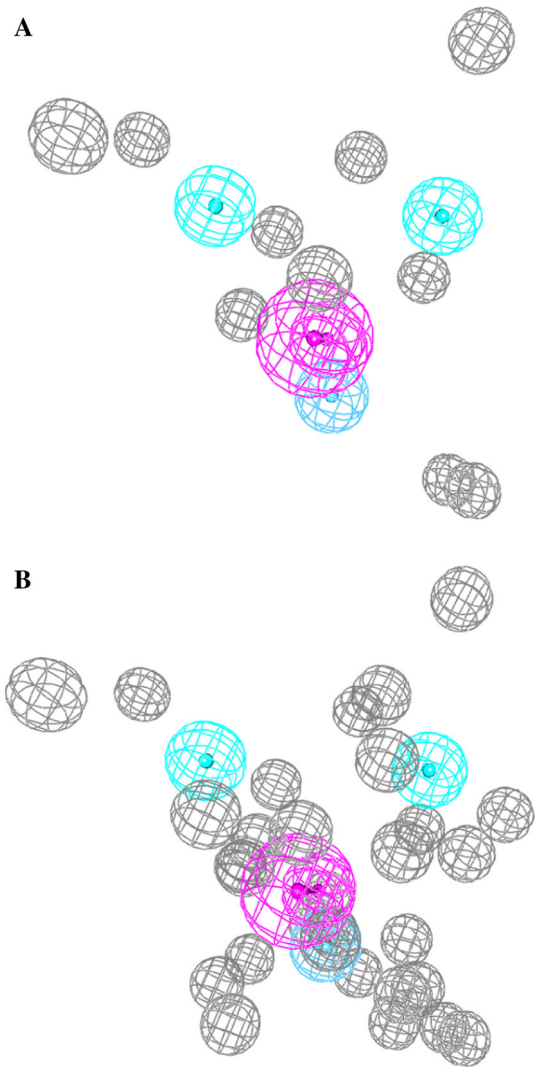


Fig. 9 **a** HypoB-I pharmacophoric features and exclusion spheres *light blue spheres* represent hydrophobic features, *dark blue spheres* represent hydrophobic aromatic feature, *vectored pink spheres* represent hydrogen bond donor, and *gray spheres* represent exclusion regions, **b** sterically-refined HypoB-I with 20 added exclusion spheres

It remains to be mentioned that dbCICA contacts summation models are judged based on three success criteria: correlation coefficient (r^2), leave-one-out r^2 (r^2LOO), or K-fold r^2 (r^2K-Fold). See "Genetic algorithm implementation in dbCICA modeling" section for more details.

Inferences from dbCICA output

From the above, one can summarize information collected from dbCICA modeling in the following points:

- Success in identifying a combination of ligand-receptor contact points of collinear relationship with bioactivity suggests validity of the corresponding docking/scoring approach. This is reminiscent of the use of other 3D-QSAR methods, e.g., CoMFA, to validate docking/scoring methodologies [64, 82, 83].
- Successful dbCICA models point to critical amino acids involved in ligand binding (binding hot spots). This information combined with identifying the optimal docking/scoring conditions are of great help in subsequent structure-based lead optimization efforts [39].
- Successful dbCICA models can be translated into pharmacophoric hypothesis useful for in silico screening of virtual databases to identify new hits. Although many studies have been published claiming to extract a pharmacophore starting solely from the structural analysis of one or few protein–ligand complexes, however, the extracted information cannot be termed "pharmacophore" if it is not based on proper structure–activity relationship analysis [79]. Although protein–ligand complexes can clear up some ambiguities about a binding pharmacophore, however, because such complexes only display active molecules bound to the target site, they do not bring any information regarding the requirement of each interaction to the activity [39]. This makes dbCICA handy for building pharmacophore models as it is based on 3D SAR differences between active and inactive ligands within the binding site.

First case: unionized GK activators

Compounds **1–41** (Table 1) were docked into GK (PDB code: 1V4S) using LigandFit docking engine [39, 91]. Subsequently, high-ranking docking solutions were scored by 6 different scoring functions implemented within LigandFit. The cycle of docking, scoring, and dbCICA modeling was repeated to cover docking in the presence or absence of crystallographically explicit water molecules within the binding site. The docked ligands were unionizable and therefore were docked without ionization.

Subsequently, two distance thresholds: 2.5 and 3.5 Å were used to determine intermolecular ligand-binding site contacts. Accordingly, two corresponding contact binary matrices were generated for each docking solution (see "Genetic algorithm implementation in dbCICA modeling" section under Experimental). Afterwards, GA-based search was implemented to search for the best summation of ligand-receptor intermolecular contacts capable of explaining bioactivity variation across the training compounds (**1–41**, Table 1). GA was instructed to scan combinations of 2–10 directly-proportional intermolecular (positive) contacts followed by 2, 5 or 10 inversely proportional (negative) contacts.

Table 2 shows the contacts distance thresholds, number of positive and negative contacts, and statistical criteria of

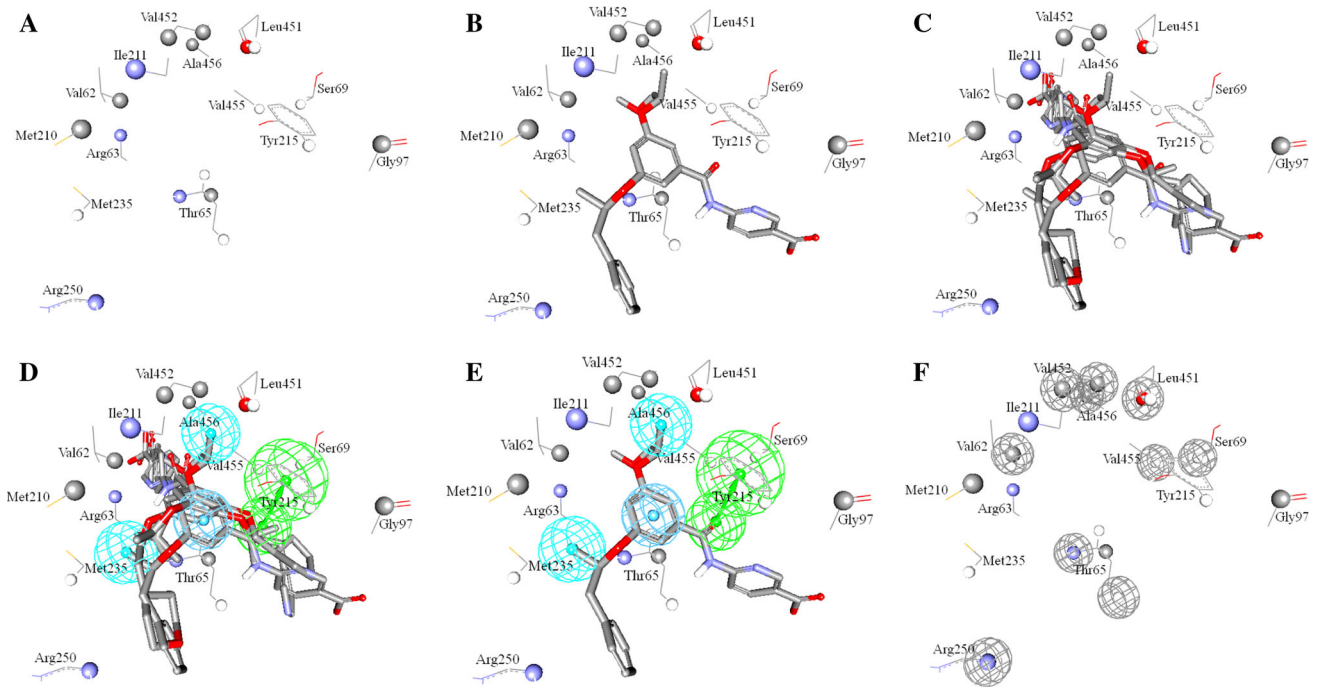


Fig. 10 Steps for manual generation of binding hypothesis HypoB-II as guided by dbCICA model B-II (Tables 13, 14): **a** the binding site moieties in dbCICA model B-II with significant contact atoms shown as *spheres*. **b** The docked pose of the well-behaved compound **71** ($EC_{50} = 0.03$ μM) within the binding pocket, **c** the docked poses of the well-behaved and potent compounds **55, 56, 62 and 71**, **d** manually placed pharmacophoric features onto chemical moieties common among docked well-behaved potent compounds **55, 56, 62 and 71**, **e** the docked pose of **71** and how it relates to the proposed pharmacophoric features. **f** Exclusion spheres fitted against binding site atoms showing negative correlations with bioactivity (as emergent in dbCICA model B-II)

optimal dbCICA models based on different docking conditions, while Table 3 summarizes the best results of the Table 2. Clearly from Table 2, all docking configurations yielded good dbCICA models with average $r^2_{5\text{-fold}}$ values of 0.60. However, docking experiments based on anhydrous binding site and via Ligscore1, PLP2 and PMF scoring functions gave the highest $r^2_{5\text{-fold}}$ (Tables 2, 3). Table 4 shows that the critical amino acid contacts in high-ranking dbCICA models.

Figures 1, 3 and 5 show how dbCICA models A-I, A-II and A-III were translated into corresponding pharmacophoric models (HypoA-1, HypoA-II and HypoA-III, respectively) employing DS 2.0 environment (see Generation of pharmacophores corresponding to successful dbCICA models section under Experimental). Initially, the binding pockets were annotated by rendering significant contacts atoms in spherical forms (see Figs. 1a, 3a, 5a). Subsequently, we selected few potent ($EC_{50} \leq 6.5$ μM) and well-behaved docked compounds, i.e., of least difference between experimental and fitted bioactivities (determined by regressing the activators' bioactivities and their docking-based contacts summations) and aligned them within the binding pocket. Thereafter, appropriate pharmacophoric features were placed onto common chemical functionalities among aligned docked compounds, as in Figs. 1d, 3d and 5d.

The following description illustrates, as an example, how HypoA-II pharmacophore model was generated from corresponding optimal dbCICA model A-II (Fig. 3): emergence of significant contact at Tyr214 (carbon CD1), combined with the consensus of well-behaved, potent docked ligands (**2, 5, 8** and **26**) on placing nearby heterocyclic rings, prompted us to place a hydrophobic feature on these rings (Fig. 3d). Similarly, agreement of docked potent well-behaved compounds on placing hydrophobic methylsulfonyl near to Ala456 (hydrogen HB1 and hydrogen HB3), prompted us to place a second hydrophobic feature onto the methyl carbon of methylsufonyl substituents (Fig. 3d). A similar conclusion was drawn from the consensus of potent well-behaved ligands on placing short hydrophobic fragments adjacent to Ile211 (hydrogen HD11 and hydrogen HA), i.e., a hydrophobic feature was placed onto those fragments. Finally, emergence of significant contact at the heterocycle rings nitrogen atoms of potent ligand with Thr65 amidic NH group, combined with the consensus of well-behaved, potent docked ligands in placing nearby amidic NH, prompted us to place a hydrogen bond acceptor on nitrogen of heterocycle rings (Fig. 3d).

A similar strategy was implemented for the development of pharmacophore models HypoA-I and HypoA-III (Figs. 1, 5, respectively) guided by dbCICA models A-I and A-III (Tables 2, 3, 4). Figures 2, 4 and 6 show the

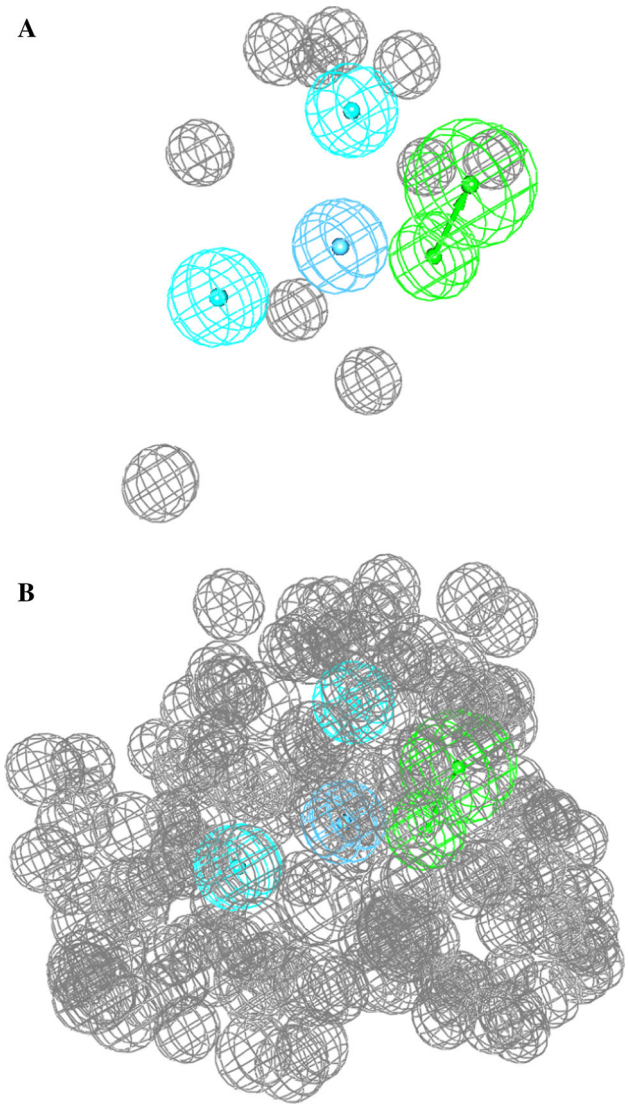


Fig. 11 **a** HypoB-II pharmacophoric features and exclusion spheres *light blue spheres* represent hydrophobic features, *dark blue spheres* represent hydrophobic aromatic features, *vectored green spheres* represent hydrogen bond acceptor features and gray spheres represent exclusion regions, **b** sterically-refined HypoB-II with 149 added exclusion spheres

resulting pharmacophore models, while Table 5 shows the X, Y, Z coordinates of the generated pharmacophores.

Validation of dbCICA models A-I, HypoA-II and HypoA-III together with corresponding HypoA-I, HypoA-II and HypoA-III

Two validation procedures were implemented to assess the dbCICA models and corresponding pharmacophores, namely: CoMFA and ROC analyses.

1. CoMFA modeling: We assessed whether the corresponding molecular alignments of dbCICA models HypoA-I, HypoA-II and HypoA-III, yield self-consistent and predictive CoMFA models [82–85]. For comparison purposes, a similar assessment was performed for molecular alignment corresponding to the lowest-ranking dbCICA model. Table 8 shows the statistical criteria of the resulting CoMFA models.

Clearly from Table 8, the statistical criteria of the top 3 dbCICA models correlate nicely with those of their CoMFA counterparts. In fact, the docking conditions corresponding to the three top dbCICA models yielded self-consistent and predictive CoMFA models (based on r^2_{PRESS}) after removal of one outlier from the testing set of each model.

On the other hand, low-ranking dbCICA model coincided with low quality CoMFA model as in Table 8: The statistical criteria of CoMFA model generated based on docking conditions of lowest ranking dbCICA failed in r^2_{PRESS} validation.

Agreement between CoMFA and dbCICA provides additional evidence on the capacity of dbCICA modeling as validation technique in docking studies.

2. Receiver operating characteristic (ROC) curve analysis generated hypothesis.

To further validate dbCICA models A-1, A-II and A-III, we decided to test the abilities of their corresponding pharmacophores, i.e., HypoA-1, HypoA-II and HypoA-III, to correctly classify a large list of virtual compounds into actives and inactives, and plot the results in ROC curves. The testing list was built by incorporating known 14 actives within 471 structurally-related decoys, as detailed in "Receiver operating characteristic (ROC) curve analysis" section under Experimental.

Figure 7a, c, e and Table 7 show the ROC performances of dbCICA-based pharmacophores. Clearly from the table and figures the three models illustrated good overall performances with ROC–AUC values exceeding 84 %.

Nevertheless, in order to further optimize the classification powers of the pharmacophores, we decorated them with steric constrains (exclusion spheres) to represent sterically forbidden regions within the binding pocket of GK. This was performed using HipHop-Refine module within DS 2.5. Table 6 lists the compounds used for steric refinement and their corresponding HipHop parameters.

Clearly from Fig. 7b, d, f and Table 7, that steric refinement enhanced the classification power of the three pharmacophores as their corresponding ROC–AUC were significantly higher than their unrefined counterparts. These results are not unexpected as steric refinement enforces stringent size limitations on captured hits, such that only small molecules than can fit into the binding cavity can be captured.

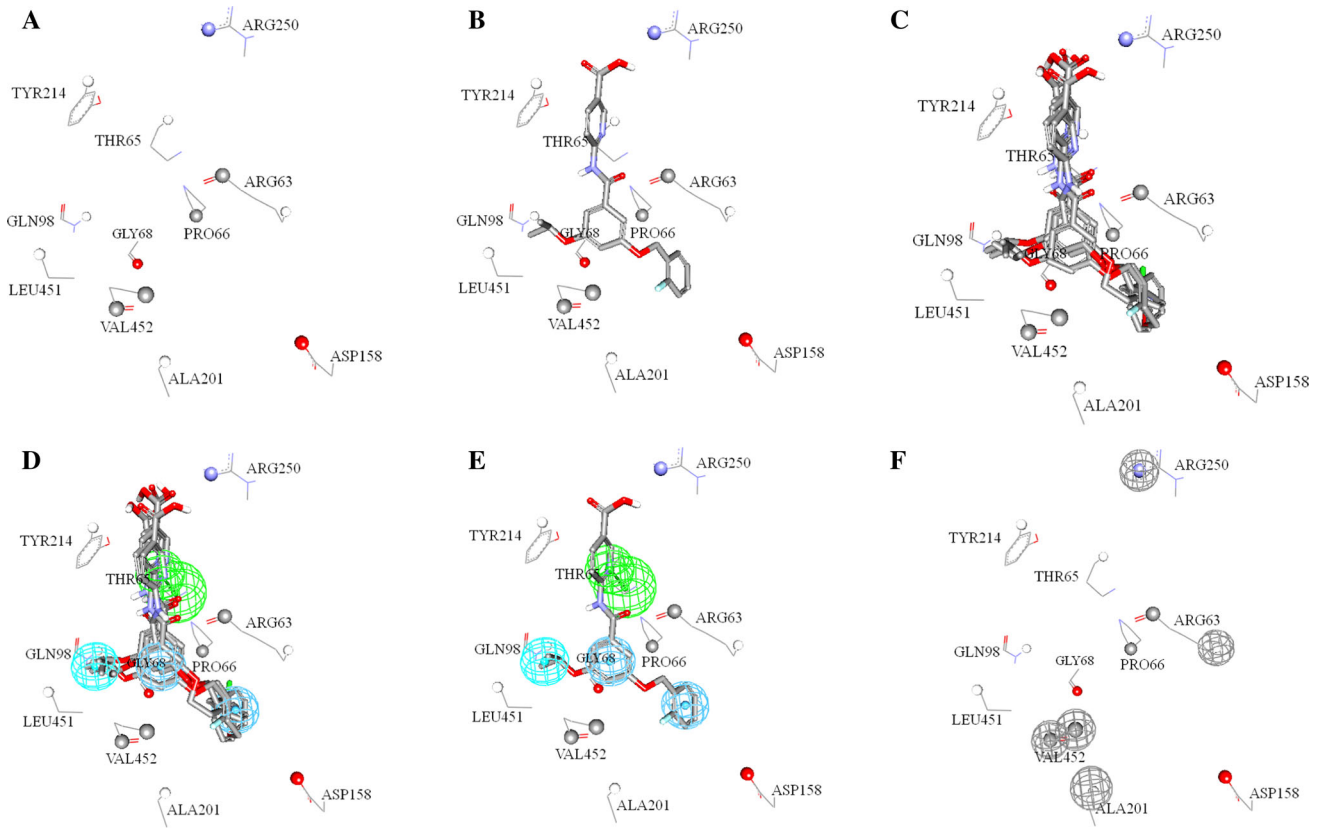


Fig. 12 Steps for manual generation of binding hypothesis HypoB-III as guided by dbCICA model B-III (Tables 13, 14): **a** the binding site moieties in dbCICA model B-III with significant contact atoms shown as *spheres*. **b** The docked pose of the well-behaved compound **64** ($EC_{50} = 0.10$ μM) within the binding pocket, **c** the docked poses of the well-behaved and potent compounds **50, 55, 58, 59** and **64**. **d** Manually placed pharmacophoric features onto chemical moieties common among docked well-behaved potent compounds **50, 55, 58, 59** and **64**. **e** The docked pose of **64** and how it relates to the proposed pharmacophoric features. **f** Exclusion spheres fitted against binding site atoms showing negative correlations with bioactivity (as emergent in dbCICA model B-III)

In silico screening and in vitro validation

In silico screening was conducted employing sterically-refined versions HypoA-I, HypoA-II and HypoA-III as 3D search queries against the National Cancer Institute list of compounds (NCI, includes 238,819 compounds) and our *in-house* list of drugs and agrochemicals (3,002 compounds). Hits were subsequently filtered based on a molecular weight threshold of 500 Da in order to remove large, non-drug-like compounds, while drugs and agrochemicals were bioassayed without any post screen filtering. The resulting hits were docked into GK protein (PDB code: 1V4S) employing the same docking conditions of corresponding dbCICA models A-I, A-II and A-III (i.e., LigandFit docking of unionized ligands into anhydrous binding pocket employing Ligscore1, PLP2 and PMF scoring functions, respectively, as in Table 2). The resulting docked poses were subsequently analyzed for critical contacts according to dbCICA models I, II and III (Table 3) and the sums of critical contacts for each hit compound were used to rank hits and prioritize subsequent in vitro testing (Table 8).

Tables 9, 10 and 11 show the highest predicted hits captured by HypoA-I, HypoA-II and HypoA-III, respectively, and their experimental in vitro bioactivities. The results further assure the validity of dbCICA based models and their ability to capture diverse active hits with excellent bioactivities. The most potent hits were **75**, **90**, **93**, **117** and **118** with GK bioactivation exceeding 7.0-folds compared to unactivated enzyme control.

Second case study: ionizable GK activators

All 30 ionizable GKAs (**42–71**, Table 1) were docked into GKA binding site employing LigandFit [37, 91]. In each case, ionized and unionized ligands were docked into hydrous and anhydrous versions of the binding pocket. High-ranking docked poses were scored by the same six different scoring functions implemented within LigandFit. The cycle of docking, scoring, and dbCICA modeling was

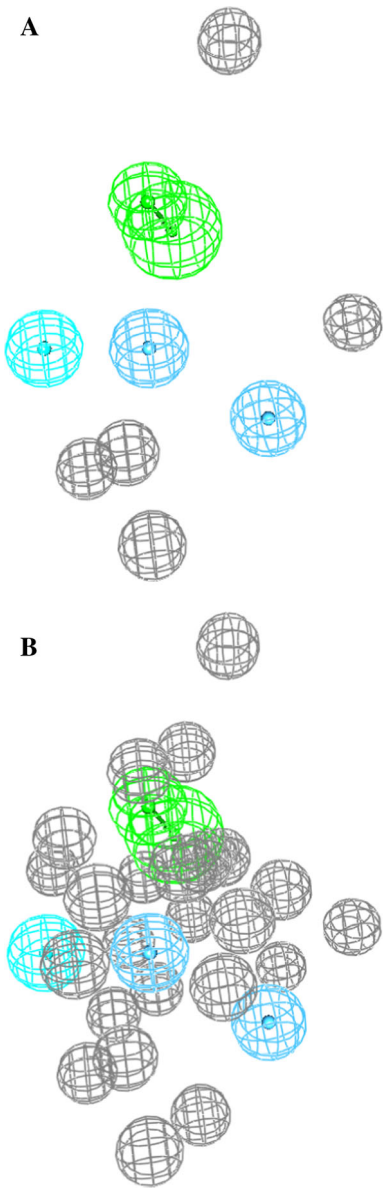


Fig. 13 **a** HypoB-III pharmacophoric features and exclusion spheres *light blue spheres* represent hydrophobic features, *dark blue spheres* represent hydrophobic aromatic features, *vectored green spheres* represent hydrogen bond acceptor features and *gray spheres* represent exclusion regions, **b** sterically-refined HypoB-III with 20 added exclusion spheres

repeated to cover all possible docking combinations resulting from the presence (or absence) of crystallographically explicit water molecules within the binding site and the ionized, un-ionized states of the ligands. The same distance thresholds and genetic algorithm were used as in the first case.

Table 12 shows the contacts distance thresholds, number of positive and negative contacts, and statistical criteria of optimal second case dbCICA models, while Table 13 summarizes the best results of Table 12. Clearly from Table 12, the resulting dbCICA models were fairly good with $r^2_{5\text{-fold}}$ values $\geq$0.73 regardless to the implemented scoring functions, ligands' ionization state or binding site hydration condition. Nevertheless, three dbCICA models significantly outperformed others, namely, B-I, B-II and B-III (Table 13).

Figures 8, 10 and 12 show how dbCICA models B-I, B-II and B-III were translated into corresponding pharmacophoric models (HypoB-I, HypoB-II and HypoB-III, respectively) employing DS 2.0 environment (see Generation of pharmacophores corresponding to successful dbCICA models section). The pharmacophoric features were placed in such away to highlight the interactions encoded by nearest critical contacts. Table 14 shows critical amino acid contacts emerged in best dbCICA models.

The following shows how pharmacophore model HypoB-I, as an example, was generated from the corresponding dbCICA model B-I (Fig. 8): Emergence of significant contacts at Ile159 (hydrogen HG21), Tyr61 (carbon C) and Val62 (hydrogen HA), combined with the consensus of well-behaved, potent docked ligands (**60, 61, 62, 69, 70** and **71**) on placing nearby pyridine rings, prompted us to place a hydrophobic aromatic on this pyridine ring (Fig. 8d). Similarly, agreement among these docked potent well-behaved compounds on placing hydrophobic aliphatic moieties near to Met210 (carbon CE), Met235 (hydrogen HE3 and sulfur SD), prompted us to place a hydrophobic feature onto these aliphatic side chains (Fig. 8d). A similar conclusion was drawn from the consensus of the same ligands on placing short hydrophobic fragments adjacent to Glu67 (hydrogen HA) and Tyr215 (carbon, CD1), i.e., a hydrophobic feature was placed onto those fragments. Emergence of significant contact at the Arg63 (oxygen O), combined with the consensus of well-behaved potent docked ligands on placing nearby amidic NH, prompted us to place a hydrogen bond donor on this NH (Fig. 8d).

A similar strategy was implemented for the development of pharmacophore models HypoB-II and HypoB-III (Figs. 10, 12, respectively) guided by dbCICA model B-II and B-III (Tables 12, 13, 14). Figures 9, 11 and 13 show the resulting pharmacophore models, while Table 15 shows the X, Y, Z coordinates of the generated pharmacophores.

Validation of dbCICA models B-I, B-II and B-III together with corresponding HypoB-I, HypoB-II and HypoB-III

Similarly, CoMFA and ROC were implemented to validate the dbCICA models and corresponding pharmacophores.

1. CoMFA modeling: To validate optimal dbCICA models (models B-I, B-II and B-III, Tables 12, 13),

Table 15 Pharmacophoric features, corresponding tolerances and 3D coordinates (X, Y, Z) of optimal dbCICA-based pharmacophore models generated from compounds **(42–71,** Table 1**)**

Model[a]	Definitions		Chemical features				
			HBD[c]		Hbic[d]	HbicArom[e]	Hbic
HypoB-I[g]	Tolerances[b]		1.6 (tail)	2.2 (head)	1.6	1.6	1.6
	Coordinates	X	38.97	41.21	41.36	36.90	43.72
		Y	17.49	16.38	11.03	18.90	19.51
		Z	60.06	58.40	62.09	58.86	64.71
Model[a]	Definitions		Chemical features				
			HbicArom	Hbic	Hbic	HBA[f]	
HypoB-II[h]	Tolerances		1.6	1.6	1.6	1.6 (tail)	2.2 (head)
	Coordinates	X	42.07	41.57	38.19	43.91	42.46
		Y	14.91	11.35	17.06	17.97	18.31
		Z	63.96	61.57	65.62	64.73	67.34
Model[a]	Definitions		Chemical features				
			Hbic	HbicArom	HbicArom	HBA	
HypoB-III[i]	Tolerances		1.6	1.6	1.6	1.6 (tail)	2.2 (head)
	Coordinates	X	39.41	38.30	36.20	42.92	43.74
		Y	17.97	17.26	20.45	13.60	16.24
		Z	66.35	62.11	57.33	61.49	60.32

[a] Pharmacophoric hypothesis shown in Figs. 9, 11 and 13

[b] Tolerances: refer to the radius of feature spheres (Å)

[c] *HBD* Hydrogen bond donor feature

[d] *Hbic* Hydrophobic feature

[e] *HbicArom* Hydrophobic aromatic feature

[f] *HBA* Hydrogen bond acceptor feature

[g] Number of exclusion spheres in HypoB-I before HipHop-steric refinement = 10 of 12 Å tolerance, at the following X,Y,Z coordinates: (44.51, 5.98,59.49), (49.02, 17.49, 68.58), (38.49, 3.99, 64.67), (36.59,26.31, 7.46), (34.69, 26.17, 58.41), (39.38, 68.7, 65.48), (43.01, 15.51, 59.65), (41.55,16.05, 68.31), (39.55, 21.03, 64.99), and (36.46, 14.06, 61.08)

[h] Number of exclusion spheres in HypoB-II before HipHop-steric refinement = 10 of 12 Å tolerance, at the following X,Y,Z coordinates: (33.76, 22.27, 61.36), (45.76, 8.43, 57.93), (37.49, 19.72, 66.78), (41.79, 22.83, 66.00), (43.01, 15.51, 59.65), (46.56, 15.31, 61.39), (35.35, 18.55, 65.44), (35.17, 16.04, 65.07), (40.59, 21.82, 63.78), and (36.46, 14.06, 61.08)

[i] Number of exclusion spheres in HypoB-III before HipHop-steric refinement = 5 of 12 Å tolerance, at the following X,Y,Z coordinates: (30.75, 18.59, 63.04), (45.76, 8.43, 57.93), (38.56, 17.53, 53.48), (34.70, 19.91, 65.13), and (33.97, 17.61, 63.54)

Table 16 The statistical results of the CoMFA models obtained via dbCICA-based docking/scoring combinations (in Table 12)

dbCICA models[a]	Docking conditions[a]	Scoring function[a]	Terms[b]	PC[c]	Statistical criteria			
					r^2[d]	r^2_{LOO}[e]	r^2_{BS}[f]	r^2_{PRESS}[g]
Highest[h] ranking								
B-I[h]	Ionized ligands-anhydrous binding pocket	Jain	6	3	0.939	0.574	0.885	0.764
B-II	Ionized ligands-anhydrous binding pocket	PMF	8	3	0.986	0.823	0.979	0.717
B-III	Unionized ligands-hydrous binding pocket	PLP2	8	4	0.973	0.918	0.968	0.891
Lowest ranking[i]	Ionized ligands-hydrous binding pocket	Ligscore2	7	5	0.971	0.914	0.972	0.637

[a] dbCICA models and corresponding docking-scoring conditions were selected from Tables 12 and 13

[b] Number of CoMFA descriptors in the best 3D-QSAR model

[c] Number of principal components (latent variables) in the best 3D-QSAR model

[d] Non-crossvalidated correlation coefficient for 24 training compounds

[e] Crossvalidation correlation coefficients determined by the leave-one out technique

[f] Bootstrapping correlation coefficient

[g] Predictive r^2 determined for the 6 test compounds

[h] Docking-scoring conditions of highest ranking dbCICA models. Models numbered as in Table 13

[i] Docking-scoring conditions of lowest ranking dbCICA models selected from Table 12

Table 17 Performance of dbCICA-selected pharmacophores as 3D search queries

Pharmacophore model	ROC[a]–AUC[b]	ACC[c]	SPC[d]	TPR[e]	FNR[f]
HypoB-I	91	96	98	57	20
HypoB-II	74	96	99	29	10
HypoB-III	88	96	99	29	10
Refined-HypoB-I	99	96	99	21	10
Refined-HypoB-II	86	96	99	29	10
Refined-HypoB-III	97	96	99	21	10

[a] *ROC* receiver operating characteristic

[b] *AUC* area under the curve (%)

[c] *ACC* overall accuracy (%)

[d] *SPC* overall specifity (%)

[e] *TPR* Overall true positive rate

[f] *FNR* Overall false negative rate (%)

we assessed their corresponding molecular alignments to see if they yield self-consistent and predictive CoMFA models or not [64, 82, 83]. For comparison purposes, a similar assessment was performed for molecular alignments corresponding to the lowest-ranking dbCICA model. Table 16 shows the statistical criteria of the resulting CoMFA models.

Clearly from Table 16, the statistical criteria of the top 3 dbCICA models correlate nicely with those of their CoMFA counterparts. In fact, the docking conditions corresponding to top three dbCICA models yielded self-consistent and predictive CoMFA models (based on r^2_{LOO} and r^2_{PRESS}).

On the other hand, low-ranking dbCICA models coincided with lower quality CoMFA model. Still, interestingly, the lowest ranking dbCICA model maintained acceptable statistical parameters. This could be explained by the excellent statistical qualities of all dbCICA models in Table 12. Nevertheless, the r^2_{PRESS} value of the low ranking dbCICA model (Table 16) is inferior to corresponding values of HypoB-I, HypoB-II and HypoB-III, which can be considered as additional validation of our dbCICA modeling.

2. ROC Curve Analysis Generated hypothesis.

Figure 14 and Table 17 show the ROC performances of generated dbCICA-based pharmacophores. Clearly from both table and figure that HypoB-I and HypoB-III are superior to HypoB-II, since AUC values of HypoB-II and sterically-refined version of HypoB-II are much less than those of HypoB-I and HypoB-III. Accordingly, sterically refined HypoB-I and HypoB-III were selected as 3D search queries against the NCI structural database.

In silico screening and in vitro validation

Sterically-refined HypoB-I and HypoB-III were employed as 3D search query to screen the NCI (includes 238,819 compounds). Hits were subsequently filtered based on a molecular weight threshold of 500 Da in order to remove large, non-drug-like compounds. The resulting hits were docked into GK protein (PDB code: 1V4S) employing the same docking conditions of dbCICA models B-I and B-III. The resulting docked poses were subsequently analyzed for critical contacts according to B-I and B-III (Table 14) and the sums of critical contacts for each hit compound were used to rank hits and the highest ranking hits were selected for bioassay. Tables 18 and 19 show the highest ranking hits, as well as their experimental in vitro bioactivities.

Overall, hits captured by the second case models (HypoB-I, HypoB-II, and HypoB-III) are inferior vis-à-vis their bioactivities compared to those captured by the first series (HypoA-I, HypoA-II, and HypoA-III) with their most potent hits showing GK activation of 4.0–6.0-folds (i.e., hits **133**, **134**, **135**, **141**, **145**, **147**, **148**, **150**, and **151,** as in Tables 18, 19) compared to GK bioactivation exceeding 7.0-folds with potent hits captured by the first series models.

We believe this difference is related to the fact that the first group of training compounds (**1–41**, Table 1) are structurally more diverse compared to the second group (**42–71**, Table 1), which is expected to render their

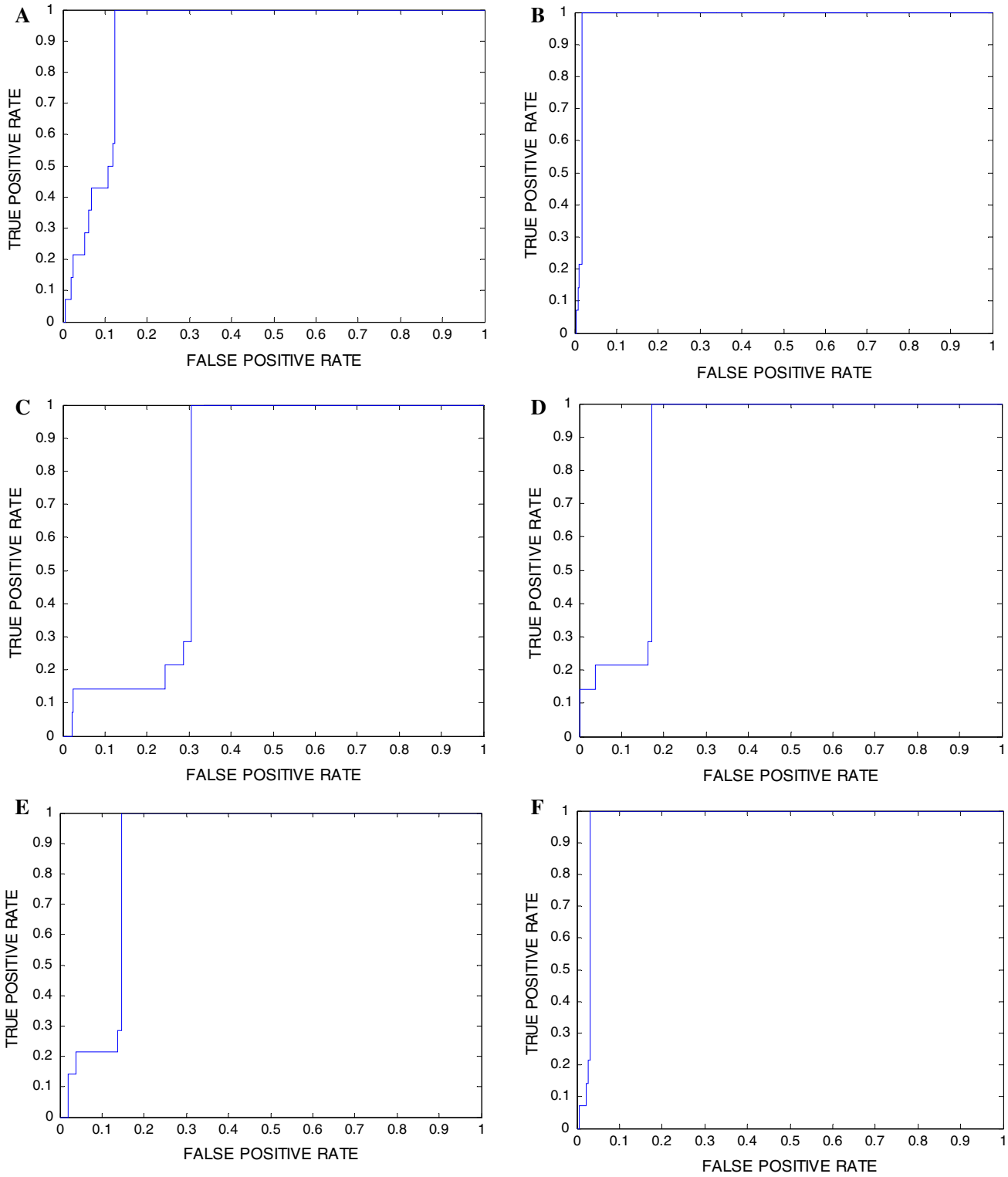
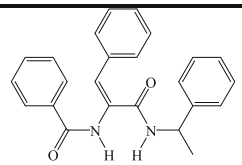
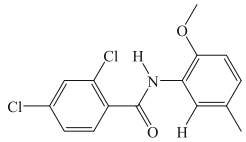
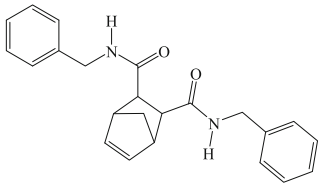
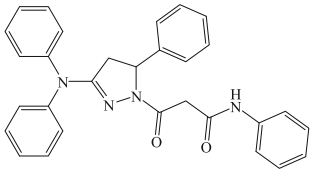
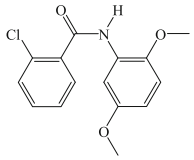
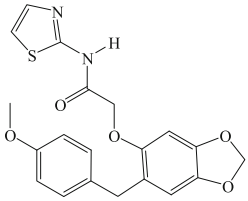
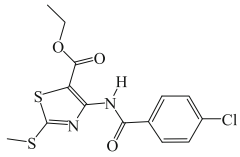
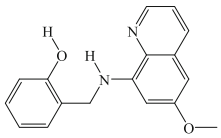
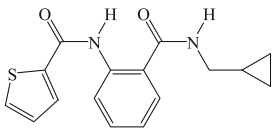
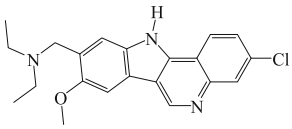


Fig. 14 Receiver operating characteristic (ROC) curves of dbCICA-based **a** HypoB-I, **b** sterically refined HypoB-I, **c** HypoB-II, **d** sterically refined HypoB-II, **e** HypoB-III and **f** sterically refined HypoB-III

Table 18 In silico hits with their fit values against (HypoB-I) and their in vitro GK activation folds

No.	NCI code or Chemical Name	Structure	dbCICA contacts[a]	Experimental Fold Activation at 10 μM
130	117590		19	3
131	204547		14	1
132	241418		19	0
133	686237		19	4
134	205524		17	4.7[b]
135	350118		20	5
136	675708		18	1
137	130801		20	1
138	379411		18	1
139	317605		16	1

[a] Contacts summations according to dbCICA model B-I (Tables 12, 13)

[b] Activation folds values were measured as duplicated readings for this hit, bioactivation values were measured only once for other hits

Table 19 In silico hits with their fit values against (HypoB-III), and their in vitro GK activation folds

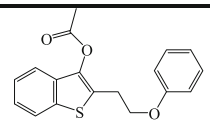
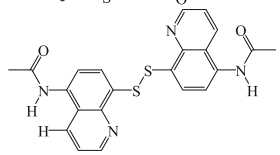
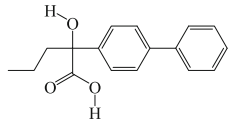
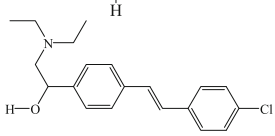
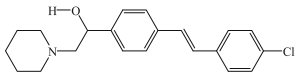
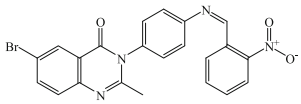
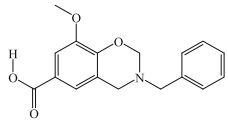
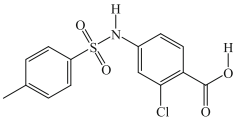
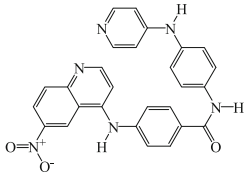
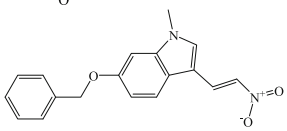
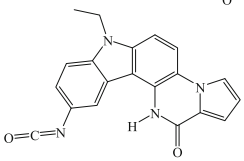
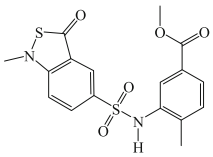
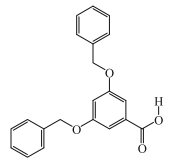
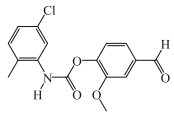
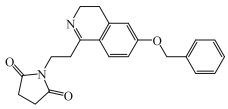
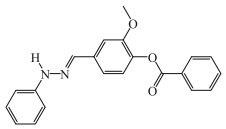
No.	NCI code or Chemical Name	Structure	dbCICA contacts[a]	Experimental Fold Activation at 10 μM
140	12029		14	0
141	14226		13	4
142	16326		15	0
143	26038		14	3
144	26040		13	1
145	90890		14	5
146	98795		18	0
147	145981		12	5
148	146771		14	6[b]
149	294438		15	0
150	340558		13	4
151	343540		14	4

Table 19 continued

No.	NCI code or Chemical Name	Structure	dbCICA contacts[a]	Experimental Fold Activation at 10 μM
152	210283		17	1
153	205638		14	0
154	153544		13	0
155	201857		15	1

[a] Contacts summations according to dbCICA model B-III (Tables 12, 13)

[b] Activation folds values were measured as duplicated readings for this hit, bioactivation values were measured only once for other hits

respective dbCICA and pharmacophore models more efficient in capturing GK bioactivators. Nevertheless, emergence of multiple successful dbCICA models probably corresponds to multiple binding modes assumed by ligands with GK binding site.

Conclusions

In the current project, we implemented a wide range of docking configurations to dock 71 activators into the binding pocket of GK. The ligands were classified into two sets according to their ionizabilities. We employed our novel dbCICA methodology to identify and validate optimal docking configurations. The resulting dbCICA models were also used to construct corresponding pharmacophore models that were validated by ROC analysis. The best pharmacophores were used to screen the NCI list of compounds. Several hits exhibited significant GK bioactivation, with the most potent hit illustrating 7.5-folds GK activation at 10 μM.

References

1. Brocklehurst JK, Payne AV, Davies AR, Carroll D, Vertigan LH, Wightman JH, Aiston S, Waddell DI, Leighton B, Coghlan PM, Agius L (2004) Stimulation of hepatocyte glucose metabolism by novel small molecule glucokinase activators. Diabetes 53:535–541
2. Leighton B, Atkinson A, Coghlan PM (2005) Small molecule glucokinase activators as novel anti-diabetic agents. Biochem Soc Trans 33:371–374
3. Kietzmann T, Ganjam KG (2005) Glucokinase: old enzyme, new target. Expert Opin Ther Pat 15:705–713
4. Sarabu R, Taub R, Grimsby J (2007) Glucokinase activation—a strategy for T2D therapy: recent developments. Drug Discov Today Ther Strateg Drug 4:111–115ti
5. Heuser S, Barrett GD, Berg M, Bonnier B, Kahl A, Puente LM, Oram N, Ried R, Roettig U, Gil SG, Seger E, Steggles JD, Wannera J, Weicherta JA (2006) Synthesis of novel cyclopropylic sulfones and sulfonamides acting as glucokinase activators. Tetrahedron Lett 47:2675–2678
6. Ishikawa M, Nonoshita K, Ogino Y, Nagae Y, Tsukahara D, Hosaka H, Maruki H, Ohyama S, Yoshimoto R, Sasaki K, Nagata Y, Eiki J, Nishimura T (2009) Discovery of novel 2-(pyridine-2-yl)-1H-benzimidazole derivatives as potent glucokinase activators. Bioorg Med Chem Lett 19:4450–4454
7. Zhang L, Li H, Zhu Q, Liu J, Chen L, Leng Y, Jiang H, Liu H (2009) Benzamide derivatives as dual-action hypoglycemic agents that inhibit glycogen phosphorylase and activate glucokinase. Bioorg Med Chem 13:4385–4388
8. Nishimura T, Iino T, Mitsuya M, Bamba M, Watanabe H, Tsukahara D, Kamata K, Sasaki K, Ohyama S, Hosaka H, Futamura M, Nagata Y, Eiki J (2009) Identification of novel and potent 2-amino benzamide derivatives as allosteric glucokinase activators. Bioorg Med Chem Lett 19:1357–1360
9. Petit P, Antoine M, Ferry G, Boutin JA, Lagarde A, Gluais L, Vincentelli R, Vuillard L (2011) The active conformation of Glucokinase is not altered by allosteric activators. Acta Crystallogr D Biol Crystallogr 67:929–935

10. Takahashi K, Hashimoto N, Nakama C, Kamata K, Sasaki K, Yoshimoto R, Ohyama S, Hosaka H, Maruki H, Nagata Y, Eiki J, Nishimura T (2009) The design and optimization of a series of 2-(pyridin-2-yl)-1H-benzimidazole compounds as allosteric glucokinase activators. Bioorg Med Chem 17:7042–7051
11. Bebernitz GR, Beaulieu V, Dale BA, Deacon R, Duttaroy A, Gao J, Grondine MS, Gupta RC, Kakmak M, Kavana M, Kirman LC, Liang J, Maniara WM, Munshi S, Nadkarni SS, Schuster HF, Stams T, St Denny I, Taslimi PM, Vash B, Caplan SL (2009) Investigation of functionally liver selective glucokinase activators for the treatment of type 2 diabetes. J Med Chem 52:6142–6152
12. Tagami S, Sekine SI, Kumarevel T, Hino N, Murayama Y, Kamegamori S, Yamamoto M, Sakamoto K, Yokoyama S (2010) Crystal structure of bacterial RNA polymerase bound with a transcription inhibitor protein. Nature 468:978–982
13. Mitsuya M, Kamata K, Bamba M, Watanabe H, Sasaki Y, Sasaki K, Ohyama S, Hosaka H, Nagata Y, Eiki J, Nishimura T (2009) Discovery of novel 3,6-disubstituted 2-pyridinecarboxamide derivatives as GK activators. Bioorg Med Chem Lett 19:2718–2721
14. Diaz A, Guivovart JJ, Fita I, Ferrer JC (2011) Crystal structure of Pyrococcus abyssi glycogen synthase with open and closed conformations. Protein Databank Entry: **3FRO**
15. Verdonk ML, Berdini V, Hartshorn MJ, Mooij WTM, Murray CW, Watson P (2004) Virtual screening using protein-ligand docking: avoiding artificial enrichment. J Chem Inf Comput Sci 44:793–806
16. Kamata K, Mitsuya M, Nishimura T, Eiki J, Nagata Y (2004) Structural basis for allosteric regulation of the monomeric allosteric enzyme human glucokinase. Structure 12:429–438
17. Klebe G (2006) Virtual ligand screening: strategies, perspectives and limitations. Drug Discov Today 11:580–594
18. Steuber H, Zentgraf M, Gerlach C, Sotriffer CA, Heine A, Klebe G (2006) Expect the unexpected or caveat for drug designers: multiple structure determinations using aldose reductase crystals treated under varying conditions. J Mol Biol 363:174–187
19. Stubbs MT, Reyda S, Dullweber F, Moller M, Klebe G, Dorsch D, Mederski W, Wurziger H (2002) pH-dependent binding modes observed in trypsin crystals: lessons for structure-based drug design. ChemBioChem 3:246–249
20. DePristo MA, de Bakker PIW, Blundell TL (2004) Heterogeneity and inaccuracy in protein structures solved by X-ray crystallography. Structure 12:831–838
21. Song CM, Lim SJ, Tong JC (2009) Recent advances in computer-aided drug design. Brief Bioinform 10:579–591
22. Menikarachchi LC, Gascon JA (2010) QM/MM approaches in medicinal chemistry research. Curr Top Med Chem 10:46–54
23. Jorgensen WL (2009) Efficient drug lead discovery and optimization accounts. Chem Res 42:724–733
24. Hecht D, Fogel GB (2009) Computational intelligence methods for docking scores. Curr Comput Aided Drug 5:56–68
25. Beeley NRA, Sage C (2003) GPCRs: An update on structural approaches to drug discovery. Targets 2:19–25
26. Morris GM, Olson AJ, Goodsell DS (2000) Protein–Ligand docking methods. Princ Med Chem 8:31–48
27. Kontoyianni M, McClellan LM, Sokol GS (2004) Evaluation of docking performance: comparative data on docking algorithms. J Med Chem 47:558–565
28. Beier C, Zacharias M (2010) Tackling the challenges posed by target flexibility in drug design. Expert Opin Drug Discov 5:347–359
29. Boyd S (2007) FlexX suite. Chem World-UK 4:72
30. Rarey M, Kramer B, Lengauer T, Klebe G (1996) A fast flexible docking method using an incremental construction algorithm. J Mol Biol 261:470–489
31. Ewing TJA, Makino S, Skillman AG, Kuntz ID (2001) DOCK 40: Search strategies for automated molecular docking of flexible molecule databases. J Comput Aided Mol Des 15:411–428
32. Jones G, Willett P, Glen RC, Leach AR, Taylor R (1997) Development and validation of a genetic algorithm for flexible docking. J Mol Biol 267:727–748
33. Vaque M, Ardrevol A, Blade C, Salvado MJ, Blay M, Fernandez-Larrea J, Arola L, Pujadas G (2008) Protein-ligand docking: a review of recent advances and future perspectives. Curr Pharm Anal 4:1–19
34. Cosconati S, Forli S, Perryman AL, Harris R, Goodsell DS, Olson AJ (2010) Virtual screening with AutoDock: theory and practice. Expert Opin Drug Dis 5:597–607
35. Morris GM, Goodsell DS, Halliday RS, Huey R, Hart WE, Belew RK, Olson AJ (1998) Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function. J Comput Chem 19:1639–1662
36. Halgren TA, Murphy RB, Friesner RA, Beard HS, Frye LL, Pollard WT, Banks JL (2004) Glide: a new approach for rapid, accurate docking and scoring 2 enrichment factors in database screening. J Med Chem 47:1750–1759
37. Accelrys Inc (2000) CERIUS2 410 LigandFit user manual. San Diego, CA
38. OpenEye Scientific Software Inc (2006) FRED: Fast rigid exhaustive docking user manual. Santa Fe
39. Diller DJ, Merz KM (2001) High throughput docking for library design and library prioritization. Proteins 43:113–124
40. Rao SN, Head MS, Kulkarni A, LaLonde JM (2007) Validation studies of the site-directed docking program LibDock. J Chem Inf Model 47:2159–2171
41. Bissantz C, Folkers G, Rognan D (2000) Protein-based virtual screening of chemical databases 1 evaluation of different docking/scoring combinations. J Med Chem 43:4759–4767
42. Gao WR, Lai YL (1998) SCORE: A new empirical method for estimating the binding affinity of a protein-ligand complex. J Mol Model 4:379–394
43. Krammer A, Kirchhoff PD, Jiang X, Venkatachalam CM, Waldman M (2005) LigScore: A novel scoring function for predicting binding affinities. J Mol Graph Model 23:395–407
44. Velec HFG, Gohlke H, Klebe G (2005) Drug score-knowledge-based scoring function derived from small molecule crystal data with superior recognition rate of near-native ligand poses and better affinity prediction. J Med Chem 48:6296–6303
45. Jain AN (2006) Scoring functions for protein-ligand docking. Curr Protein Pept Sci 7:407–420
46. Rajamani R, Good AC (2007) Ranking poses in structure-based lead discovery and optimization: current trends in scoring function development. Curr Opin Drug Disc 10:308–315
47. Krovat EM, Langer T (2004) Impact of scoring functions on enrichment in docking-based virtual screening: an application study on renin inhibitors. J Chem Inf Comput Sci 44:1123–1129
48. Foloppe N, Hubbard R (2006) Towards predictive ligand design with free-energy based computational methods? Curr Med Chem 13:3583–3608
49. Englebienne P, Moitessier N (2009) Docking ligands into flexible and solvated macromolecules 4 are popular scoring functions accurate for this class of proteins? J Chem Inf Model 49:1568–1580
50. Jain AN (1996) Scoring non-covalent protein–ligand interactions: a continuous differentiable function tuned to compute binding affinities. J Comput Aided Mol Des 10:427–440
51. Böhm HJ (1998) Prediction of binding constants of protein ligands: a fast method for the prioritization of hits obtained from de novo design or 3D database search programs. J Comput Aided Mol Des 12:309–323

52. Eldridge MD, Murray CW, Auton TR, Paolini GV, Mee RP (1997) Empirical scoring functions: I the development of a fast empirical scoring function to estimate the binding affinity of ligands in receptor complexes. J Comput Aided Mol Des 11:425–445
53. Wang R, Gao Y, Lai L (1998) SCORE: A new empirical method for estimating the binding affinity of a protein-ligand complex. J Mol Model 4:379–394
54. Gehlhaar DK, Bouzida D, Rejto P (1999) Reduced dimensionality in ligand-protein structure prediction: covalent inhibitors of serine proteases and design of site-directed combinatorial libraries. In: Parrill L, Rami Reddy M (eds) Rational drug design: novel methodology and practical applications. American Chemical Society, Washington, DC, pp 292–311
55. Wang R, Lai L, Wang S (2002) Further development and of empirical scoring functions for structure-based binding validation affinity prediction. J Comput Aided Mol Des 16:11–26
56. Muegge I, Martin YC (1999) A general and fast scoring function for protein-ligand interactions: a simplified potential approach. J Med Chem 42:791–804
57. Muegge I (2000) A knowledge-based scoring function for protein-ligand interactions: probing the reference state. Perspect Drug Discov 20:99–114
58. Muegge I (2001) Effect of ligand volume correction on PMF scoring. J Comput Chem 22:418–425
59. Gohlke H, Hendlich M, Klebe G (2000) Knowledge-based scoring function to predict protein-ligand interactions. J Mol Biol 295:337–356
60. Muegge I (2006) PMF scoring revisited. J Med Chem 49:5895–5902
61. Leach AR, Shoichet BK, Peishoff CE (2006) Prediction of protein-ligand interactions Docking and scoring: successes and gaps. J Med Chem 49:5851–5855
62. Krissinel E (2009) Crystal contacts as nature's docking solutions. J Comput Chem 31:133–143
63. Steinbrecher T, Labahn A (2010) Towards accurate free energy calculations in ligand protein-binding studies. Curr Med Chem 17:767–785
64. Taha MO, AlDhamin M (2005) Effects of variable docking conditions and scoring functions on the qualities of protein aligned CoMFA models constructed from diverse h-PTP 1B inhibitors. J Med Chem 48:8016–8034
65. Tame JRH (1999) Scoring functions: a view from the bench. J Comput Aided Mol Des 13:99–108
66. Garcia-Sosa AT, Hetenyi C, Maran U (2010) Drug efficiency indices for improvement of molecular docking scoring functions. J Comput Chem 31:174–184
67. Homans SW (2007) Water, water everywhere—except where it matters. Drug Discov Today 12:534–539
68. Poornima CS, Dean PM (1995) Hydration in drug design 1 multiple hydrogen-bonding features of water molecules in mediating protein-ligand interactions. J Comput Aided Mol Des 9:500–512
69. Poornima CS, Dean PM (1995) Hydration in drug design 2 influence of local site surface shape on water binding. J Comput Aided Mol Des 9:513–520
70. Poornima CS, Dean PM (1995) Hydration in drug design 3 conserved water molecules at the ligand-binding sites of homologous proteins. J Comput Aided Mol Des 9:521–531
71. Koehler KF, Rao SN, Snyder JP (1996) Modeling drug-receptor interactions. In: Cohen NC (ed) Guidebook on molecular modeling in drug design. Academic Press, San Diego, pp 235–336
72. Pastor M, Cruciani G, Watson KA (1997) Strategy for the incorporation of water molecules present in a ligand binding site into a three-dimensional quantitative structure-activity relationship analysis. J Med Chem 40:4089–4102
73. Garcia-Sosa AT, Mancera RL, Dean PM (2003) WaterScore: a novel method for distinguishing between bound and displaceable water molecules in the crystal structure of the binding site of protein-ligand complexes. J Mol Model 9:172–182
74. Garcia-Sosa AT (2013) Hydration properties of ligands and drugs in protein binding sites: tightly-bound, bridging water molecules and their effects and consequences on molecular design strategies. J Chem Inf Model 53:1388–1405
75. Martin YC (2009) Let's not forget tautomers. J Comput Aided Mol Des 23:693–704
76. Waszkowycz B (1998) New methods for structure-based de novo drug design. In: Harvey AL (ed) Advances in drug discovery techniques. Wiley, UK, pp 150–153
77. Sutherland JJ, Nandigam RK, Erickson JA, Vieth M (2007) Lessons in molecular recognition 2 assessing and improving cross-docking accuracy. J Chem Inf Model 47:2293–2302
78. Triballeau N, Acher F, Brabet I, Pin JP, Bertrand HO (2005) Virtual screening workflow development guided by the "Receiver Operating Characteristic" curve approach application to high-throughput docking on metabotropic glutamate receptor subtype. J Med Chem 48:2534–2547
79. Triballeau N, Bertrand HO, Acher F (2006) Are you sure you have a good model? In: Langer T, Hoffmann RD (eds) Pharmacophores and pharmacophore searches. Wiley, Weinheim, pp 325–364
80. Kirchmair J, Markt P, Distinto S, Wolber G, Langer T (2008) Evaluation of the performance of 3D virtual screening protocols: rMSD comparisons, enrichment assessments, and decoy selection What can we learn from earlier mistakes? J Comput Aided Mol Des 22:213–228
81. Wang R, Lu Y, Wang S (2003) Comparative evaluation of 11 scoring functions for molecular docking. J Med Chem 46: 2287–2303
82. Abu-Hammad AM, Afifi F, Taha MO (2007) Combining docking, scoring and molecular field analyses to probe influenza neuraminidase–ligand interactions. J Mol Graph Model 26: 443–456
83. Abu-Hammad A, Zalloum WA, Zalloum H, Abu-Sheikha G, Taha MO (2009) Homology modeling of MCH1 receptor and validation by docking/scoring and protein-aligned CoMFA. Eur J Med Chem 44:2583–2596
84. Taha MO, Habash M, Al-Hadidi Z, Al-Bakri A, Younis K, Sisan S (2011) Docking-based comparative intermolecular contacts analysis as new-3D QSAR concept for validating docking studies and in silico screening: NMT and GP inhibitors as case studies. J Chem Inf Model 51:647–669
85. Al-Sha'er MA, Taha MO (2012) Application of docking-based comparative intermolecular contacts analysis to validate Hsp90α docking studies and subsequent in silico screening for inhibitors. J Mol Model 18:4843–4863
86. Castelhano LA, Dong H, Fyfe MCT, Gardner LS, Kamikozawa Y, Kurabayashi S, Nawano M, Ohashi R, Procter MJ, Qiu L, Rasamison CM, Schofield KL, Shah VK, Ueta K, Williams GM, Wittera D, Yasuda K (2005) Glucokinase-activating ureas. Bioorg Med Chem Lett 15:1501–1504
87. Bertram LS, Black D, Briner PH, Chatfield R, Cooke A, Fyfe MC, Murray PJ, Naud F, Nawano M, Procter MJ, Rakipovski G, Rasamison CM, Reynet C, Schofield KL, Shah VK, Spindler F, Taylor A, Turton R, Williams GM, Wong-Kai-In P, Yasuda K (2008) SAR, pharmacokinetics, safety, and efficacy of glucokinase activating 2-(4-sulfonylphenyl)-N-thiazol-2-ylacetamides: discovery of PSN-GK). J Med Chem 51(14):4340–4345
88. McKerrecher D, Allen JV, Bowker SS, Boyd S, Caulkett PWR, Currie GS, Davies CD, Fenwick ML, Gaskin H, Grange E, Hargreaves RB, Hayter BR, James R, Keith M, Johnson KM, Johnstone C, Jones CD, Lackie S, Rayner JW, Walker RP

(2005) Discovery, synthesis and biological evaluation of novel glucokinase activators. Bioorg Med Chem Lett 15(8):2103–2106
89. McKerrecher D, Allen JV, Caulkett PW, Donald CS, Fenwick ML, Grange E, Johnson KM, Johnstone C, Jones CD, Pike KG, Rayner JW, Walker RP (2006) Design of a potent, soluble glucokinase activator with excellent in vivo efficacy. Bioorg Med Chem Lett 16:2705–2709
90. Gasteiger J, Marsili M (1980) Iterative partial equalization of orbital electronegativity—a rapid access to atomic charges. Tetrahedron 36:3219–3228
91. Venkatachalam CM, Jiang X, Oldfield T, Waldman M (2003) LigandFit: a novel method for the shape-directed rapid docking of ligands to protein active sites. J Mol Graph Model 21: 289–307
92. Gehlhaar DK, Verkhivker GM, Rejto PA, Sherman CJ, Fogel DB, Fogel LJ, Freer ST (1995) Molecular recognition of the inhibitor AG-1343 by HIV-1 Protease: conformationally flexible docking by evolutionary programming. Chem Biol 2:317–324
93. Accelrys Inc (2009) Discovery Studio version 2.5 (DS 2.5) user manual. San Diego, CA
94. Accelrys Inc (2005) CERIUS2 QSAR users' manual. San Diego, CA
95. Accelrys Inc (1997) CERIUS2 OFF. San Diego, pp 5–109
96. Irwin JJ, Shoichet BK (2005) ZINC—A free database of commercially available compounds for virtual screening. J Chem Inf Model 45:177–182
97. Jacobsson M, Liden P, Stjernschantz E, Bostroem H, Norinder U (2003) Improving structure-based virtual screening by multivariate analysis of scoring data. J Med Chem 46:5781–5789
98. Al-masri IM, Mohammad MK, Taha MO (2008) Discovery of DPP IV inhibitors by pharmacophore modeling and QSAR analysis followed by in silico screening. Chem Med Chem 3:1763–1779
99. Taha MO, Qandil AM, Zaki DD, AlDamen MA (2005) Ligand-based assessment of factor Xa binding site flexibility via elaborate pharmacophore exploration and genetic algorithm-based QSAR modeling. Eur J Med Chem 40:701–727
100. Taha MO, Bustanji Y, Al-Bakri AG, Yousef A-M, Zalloum WA, Al-Masri IM, Atallah N (2007) Discovery of new potent human protein tyrosine phosphatase inhibitors via pharmacophore and QSAR analysis followed by in-silico screening. J Mol Graph Model 25:870–884
101. Taha MO, Atallah N, Al-Bakri AG, Paradis-Bleau C, Zalloum H, Younis K, Levesque RC (2008) Discovery of new murf inhibitors via pharmacophore modeling and QSAR analysis followed by in silico screening. Bioorg Med Chem 16:1218–1235
102. Al-Sha'er MA, Taha MO (2010) Discovery of novel CDK1 inhibitors by combining pharmacophore modeling, QSAR analysis and in silico screening followed by in vitro bioassay. Eur J Med Chem 45:4316–4330
103. Goward CR, Hartwell R, Atkinson T, Scawen MD (1986) The purification and characterization of glucokinase from the thermophile Bacillus stearothermophilus. Biochem J 15:415–420