Natural allosteric modulators and their biological targets: molecular signatures and mechanisms

Marjorie Bruder, Gina Polo and Daniela Barretto Barbosa Trivella*

Brazilian Biosciences National Laboratory (LNBio), National Centre for Research in Energy and Material (CNPEM). 13083-970 Campinas, SP-Brazil

Supporting Information

1. Dataset compilation

A set of natural allosteric modulators was constructed from the chemical databases PubChem, SciFinder and PubMed. Abstracts and titles of journal articles were searched for the keywords: "natural product" and "allosteric modulator" or "allosteric inhibitor" or "allosteric activator" or "non-competitive" or "uncompetitive" (Table S1). All unique publications from 2008 to 2018, were considered, which totaled 1,212 compounds. After manual curation, this set was constituted of 221 single natural allosteric compounds; some compounds were modulators of more than one target, constituting a total of 47 repeats. We also considered a set of 78,810 allosteric modulators (natural, synthetic and peptides), 461 allosteric experimental drugs and 19 allosteric approved drugs obtained from the AlloStericDatabase (ASD) (mdl.shsmu.edu.cn/ASD/).[1] Besides, 1,853 non-peptide inhibitors (iPPI) across 18 families of protein-protein interactions from the iPPI-Database (ippidb.cdithem.fr/)[2] and 11,452 unique compounds from the Drug Bank Database (drugbank.ca/)[3] including 2,255 approved small molecule drugs were contemplated. The dataset was completed with 208,240 compounds from the Natural Products Database UNPD (pkuxxj.pku.edu.cn/UNPD/),[4] therefore forming our final dataset of 301,056 compounds.

Table S1. Keywords used to retrieve the allosteric set and amount of papers and compounds found for each keyword.

Keywords		Pubchem	Scifinder
'allosteric modulator' AND "natural product"	Papers	7	77
	Compounds	12	97
'allosteric inhbitor' AND "natural product"	Papers	15	97
	Compounds	139	40
'allosteric activator' AND "natural product"	Papers	3	20
	Compounds	18	41
'non-competitive' AND "natural product"	Papers	52	226
	Compounds	318	554
'uncompetitive' AND "natural product"	Papers	19	74
	Compounds	69	200
Total compounds		**477**	**932**
Single compounds after manual curation		**221**	

[1] Q. Shen, G. Wang, S. Li, X. Liu, S. Lu, Z. Chen, K. Song, J. Yan, L. Geng, Z. Huang, W. Huang, G. Chen and J. Zhang, *Nucleic Acids Res.*, **2016**, *44*, D527–D535.

[2] C. M. Labbé, M. A. Kuenemann, B. Zarzycka, G. Vriend, G. A. F. Nicolaes, D. Lagorce, M. A. Miteva, B. O. Villoutreix and O. Sperandio, *Nucleic Acids Res.*, **2016**, *44*, D542–D547.

[3] D. S. Wishart, Y. D. Feunang, A. C. Guo, E. J. Lo, A. Marcu, J. R. Grant, T. Sajed, D. Johnson, C. Li, Z. Sayeeda, N. Assempour, I. Iynkkaran, Y. Liu, A. MacIejewski, N. Gale, A. Wilson, L. Chin, R. Cummings, Di. Le, A. Pon, C. Knox and M. Wilson, *Nucleic Acids Res.*, **2018**, *46*, D1074–D1082.

[4] M. Wang and N. Bandeira, *J. Proteome Res.*, **2013**, *12*, 3944–3951.

2. Chemical Space

2.1. Molecular Descriptors Calculation

From the SMILES of the 301,057 compounds, 195 topological, 47 geometrical, and 17 constitutional descriptors were calculated using the open-source Java framework Chemistry Development Kit (CDK),[5] through the "rcdk" R package.[6] A description of the molecular descriptors calculated is found in Table S2.

Table S2. Molecular descriptors considered in the exploration of the chemical space obtained from the CDK Library

	Descriptor	Type	Description
1	nSmallRings	Topological	Total number of small rings (of size 3 through 9)
2	nAromRings	Topological	Total number of small aromatic rings
3	nRingBlocks	Topological	Total number of distinct ring blocks
4	nAromBlocks	Topological	Total number of aromatically connected components
5-11	nrings3, nRings4,nRings5, nRings6, nRings7, nRings8, nRings9	Topological	Individual breakdown of 3,4,5,6,7,8, or 9 membered rings
12	tpsaEfficiency	Topological	Polar surface area expressed as a ratio to molecular size.
13	Zagreb	Topological	Sum of the squares of atom degree over all heavy atoms
14-15	WienerNumbers (WPATH,WPOL)	Topological	Wiener path number and Wiener polarity number. Sum of the lengths of the shortest paths between all pairs of vertices in the chemical graph representing the non-hydrogen atoms in the molecule.
16-20	WeightedPath (WTPT.1, WTPT.2, WTPT.3, WTPT.4, WTPT.5)	Topological	The weighted path (molecular ID) descriptors described by Randic.[7] They characterize molecular branching.
21	VadjMat	Topological	Vertex adjacency information (magnitude): $1 + \log_2(m)$, where m is the number of heavy-heavy bonds.
22	VABC	Topological	Volume of a molecule
23	TopoPSA	Topological	Calculation of topological polar surface area based on fragment contributions
24-25	TopoShape,geomShape	Topological/Geometrical	The topological and geometric shape indices described Petitjean and Bath et al, respectively.[8,9] Both measure the anisotropy in a molecule.
6	PetitjeanNumber	Topological	Petitjean Number of a molecule. According to the Petitjean definition,[8] the eccentricity of a vertex corresponds to the distance from that vertex to the most remote vertex in the graph.
27-45	MDEC.11, MDEC.12, MDEC.13, MDEC.14, MDEC.22, MDEC.23, MDEC.24, MDEC.33, MDEC.34, MDEC.44, MDEO.11, MDEO.12, MDEO.22, MDEN.11, MDEN.12, MDEN.13, MDEN.22, MDEN.23, MDEN.33	Topological	Evaluate molecular distance edge descriptors for C, N and O
46-124	KierHallSmarts (khs.sLi, khs.ssBe, khs.ssssBe, khs.ssBH, khs.sssB, khs.ssssB, khs.sCH3, khs.dCH2, khs.ssCH2, khs.tCH, khs.dsCH, khs.aaCH, khs.sssCH, khs.ddC, khs.tsC, khs.dssC, khs.aasC, khs.aaaC, khs.ssssC, khs.sNH3, khs.sNH2, khs.ssNH2, khs.dNH, khs.ssNH, khs.aaNH, khs.tN, khs.sssNH, khs.dsN, khs.aaN, khs.sssN, khs.ddsN, khs.aasN, khs.ssssN, khs.sOH, khs.dO, khs.ssO, khs.aaO, khs.sF, khs.sSiH3, khs.ssSiH, khs.ssSiH2, khs.sssSi, khs.sPH2, khs.ssssSi, khs.sPH2, khs.ssPH, khs.sssP, khs.dsssP, khs.sssssP. khs.sSH, khs.dS, khs.ssS, khs.aaS, khs.dssS,	Topological	Counts the number of occurrences of the E-state fragments[10]

[5] E. L. Willighagen, J.W. Mayfield, J. Alvarsson, A. Berg, L. Carlsson, N. Jeliazkova, S. Kuhn T. Pluskal M. Rojas-Chertó O. Spjuth, G. Torrance, C. Evelo, R Guha, C. Steinbeck, *Journal of Cheminformatics*, **2017**, *9*(1), 33.

[6] R. Guha, *Journal of Statistical Software*, **2007**, *6*,18.

[7] M. Randić, SC., Basak *Journal of Chemical Information and Computer Sciences*, **1999**, *39*(2), 261-266.

[8] M. Petitjean, *Journal of Chemical Information and Computer Sciences*, **1992**, *32*(4), 331-337.

[9] PA. Bath, AR. Poirrette, P. Willett, FH. Allen, *Journal of Chemical Information and Computer Sciences*, **1995**, *35*(4), 714-716.

[10] K. Roy, I. Mitra, *Current Computer-aided Drug Design*, **2012**, *8*(2), 135-158.

	khs.ddssS, khs.sCl, khs.sGeH3, khs.ssGeH2, khs.sssGeH, khs.ssssGe, khs.aAsH2, khs.ssAsH, khs.sssAs, khs.sssdAs, khs.ssssssAs, khs.sSeH, khs.dSe, khs.ssSe, khs.aaSe, khs.dssSe, khs.ddssSe, khs.sBr, khs.sSnH3, khs.ssSnH2, khs.sssSnH, khs.ssssSn, khs.sI, khs.PbH3, khs.PbH2, khs.sssPbH, khs.ssssPb).		
125-127	Kier1, Kier2, Kier3	Topological	Calculation of Kier and Hall kappa molecular shape indices.
128	HybRatio	Topological	Fraction of sp3 carbons to sp2 carbons.
129	fragC	Topological	Fragment complexity
130	FMF	Topological	Ratio of heavy atoms in the framework to the total number of heavy atoms in the molecule. By definition, acyclic molecules which have no frameworks, will have a value of 0.
131	ECCEN	Topological	Eccentric connectivity index that combines distance and adjacency information.[11]
132-146	ChiPath (SP.0, SP.1, SP.2, SP.3, SP.4, SP.5, SP.6, SP.7, VP.1, VP.2, VP.3, VP.4, VP.5, VP.6, VP.7)	Topological	Evaluates the Kier & Hall Chi path indices of orders 0,1,2,3,4,5,6 and 7
147-153	ChiPathCluster (VP.0, SPC.4, SPC.5, SPC.6, VPC.4, VPC.5, VPC.6)	Topological	Evaluates the Kier & Hall Chi path cluster indices of orders 4,5 and 6
154-161	ChiCluster (SC.3, SC.4, SC.5, SC.6, VC.3, VC.4, VC.5, VC.6)	Topological	Evaluates the Kier & Hall Chi cluster indices of orders 3,4,5 and 6
162-171	ChiChain (SCH.3, SCH.4, SCH.5, SCH.6, SCH.7, VCH.3, VCH.4, VCH.5, VCH.6, VCH.7)	Topological	Evaluates the Kier & Hall Chi chain indices of orders 3,4,5,6 and 7.
172-180	CarbonTypes(C1SP1, C2SP1, C1SP2, C2SP2, C3SP2, C1SP3, C2SP3, C3SP3, C4SP3)	Topological	Characterizes the carbon connectivity in terms of hybridization. (Ex. C3SP2- Doubly bound carbon bound to three other carbons)
181-185	AutocorrelationPolarizability (ATSp1, ATSp2, ATSp2, ATSp4, ATSp5)	Topological	Moreau-Broto autocorrelation using polarizability
186-190	AutocorrelationMass (ATSm1, ATSm2, ATSm3, ATSm4, ATSm5)	Topological	Moreau-Broto autocorrelation descriptors using atomic weight.
191-195	AutocorrelationCharge (ATSc1, ATSc2, ATSc3, ATSc4, ATSc5)	Topological	The Moreau-Broto autocorrelation descriptors using partial charges
196-202	MomentOfInertia (MOMI.X, MOMI.Y, MOMI.Z, MOMI.XY, MOMI.XZ, MOMI.YZ, MOMI.R)	Geometrical	Principal moments of inertia and ratios of the principal moments. Also calculates the radius of gyration.
203-204	LengthOverBreadth (LOBMAX, LOBMIN)	Geometrical	Calculates the ratio of length to breadth.
205-213	GravitationalIndex (GRAV.1, GRAV.2, GRAV.3, GRAVH.1, GRAVH.2, GRAVH.3. GRAV.4, GRAV.5, GRAV.6)	Geometrical	Mass distribution of the molecule.
214-242	CPSA (PPSA.1, PPSA.2, PPSA.3, PNSA.1, PNSA.2, PNSA.3, DPSA.1, DPSA.2, DPSA.3, FPSA.1, FPSA.2, FPSA.3, FNSA.1, FNSA.2, FNSA.3, WPSA.1, WPSA.2, WPSA.3, WNSA.1, WNSA.2, WNSA.3, RPCG, RNCG, RPC5, RNCS, THSA, TPSA, RHSA, RPSA)	Geometrical	A variety of descriptors combining surface area and partial charge information. Capture information about the features of molecules responsible for polar intermolecular interactions.[12]
243	XLogP	Constitutional	Prediction of logP based on the atom-type method called XLogP.
244	Weight (MW)	Constitutional	Weight of atoms of a certain element type. If no element is specified, the returned value is the Molecular Weight
245	Lipinski Failures	Constitutional	Number failures of the Lipinski's Rule of Five.
246	RotatableBondsCount (nRotB)	Constitutional	Number of nonrotatable bonds on a molecule.
247	MLogP	Constitutional	LogP based on a the Mannhold equation using the number of carbons and hetero atoms.
248	LongestAliphaticChain (nAtomLAC)	Constitutional	Number of atoms in the longest aliphatic chain.
249	LargestPiSystem (nAtomP)	Constitutional	Number of atoms in the largest pi system.
250	LargestChain (nAtomLC)	Constitutional	Number of atoms in the largest chain
251	BondCount (nB)	Constitutional	Number of bonds of a certain bond order.
252	nBase	Constitutional	Returns the number of basic groups.

[11] V. Sharma, R. Goswami, AK., Madan *Journal of Chemical Information and Computer Sciences*,**1997**, *37*(2), 273-282.

[12] DT. Stanton, S. Dimitrov, V. Grancharov, OG. Mekenyan *SAR and QSAR in Environmental Research*, **2002**, *13*(2), 341-351.

253	AtomCount (nAtom)	Constitutional	Number of atoms of a certain element type.
254	AromaticBondsCount (nAromBond)	Constitutional	Number of aromatic bonds of a molecule.
255	AromaticAtomsCount (naAromAtom)	Constitutional	Number of aromatic atoms of a molecule.
256-258	AlogP, Alogp2, AMR	Constitutional	Atom additive logP and molar refractivity values as described by Ghose and Crippen.[13]
259	nAcid	Constitutional	Returns the number of basic groups.

A Principal Component Analysis (PCA) was performed to reduce the dimensionality of the 259-dimensional dataset and elucidate the chemical space covered by the compounds, while maintaining as much variability as possible. This is equivalent to solving an eigenvalue/eigenvector problem, since the eigenvalues are the variances of the linear combinations defined by the corresponding eigenvector.[14] In this way we found the PCs from the original (centered) dataset descriptors, that successively maximize variance and that are uncorrelated with each other.[15] As PCA may be dominated by variables with large units of measurement, we standardized the molecular descriptors by centered and divided each value x_{ij} by the standard deviation s_j of the n observations of each descriptor j,

$$Y_{ij} = \frac{x_{ij} - \grave{x}_j}{s_j}$$

(1)

Thus, the initial data matrix was replaced with the standardized matrix $\mathbf{Y_{ij}}$. Thus, since the covariance matrix of a standardized dataset is plainly a correlation matrix of the original dataset, a PCA on the standardized data is also known as a correlation matrix PCA. The covariance matrix can be defined as:

$$S = \frac{1}{N}\sum(x_n - \grave{x})(x_n - \grave{x})^T$$

(2)

where each compound x_n is projected onto a scalar value, x-bar is the sample set mean given by:

$$\grave{x} = \frac{1}{N}\sum x_n$$

(3)

In this way, the most similar molecular descriptors (represented by vectors) are grouped in each quadrant of a 2D (PC1, PC2) plane, and hence the most related compounds. PC1 represents the direction in feature space along which projections have the largest variance and PC2 is the direction which maximizes variance among all directions orthogonal to the first. Table S3 quantitatively represents the distribution of the different datasets in the chemical space created from the PCA. Accordingly, the cosine of the angle between two vectors is the coefficient of correlation between those molecular descriptors. Similarly, the cosine of the angle between any vector and the axis representing a PC is the coefficient of correlation between these features. Variables that do not correlate with any PC or correlate with the last dimensions are variables with low contribution. The corresponding variable map is illustrated in Figure S2, and statistical data regarding dataset distribution throughout the 4 quadrants is shown in Table S3.

[13] AK. Ghose, A. Pritchett, GM. Crippen, *J Comput Chem*, **1988**, *9*(1),80–90.

[14] IT. Jolliffe, J. Cadima, *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, **2016**, *374*(2065).

[15] J.L, Reymond, M. Awale, *ACS Chemical Neuroscience*, **2012**,*3*(9), 649-657.

Table S3. Quantitative distribution of the different datasets in the chemical space.

Quadrant	UNPD-ISDB	ASD (non-drugs)	iPPIs (non-peptides)	DrugBank (experimental)	DrugBank (approved)	ASD (exp. drugs)	ASD (app. drugs)	alloNPs	Total
Q1	78398	6674	92	3947	1068	28	2	40	90249
Q2	33810	58726	998	4081	826	414	12	106	98973
Q3	69406	513	16	433	173	6	2	40	70589
Q4	26626	12897	748	736	188	13	3	35	41246
TOTAL	208240	78810	1854	9197	2255	461	19	221	301057
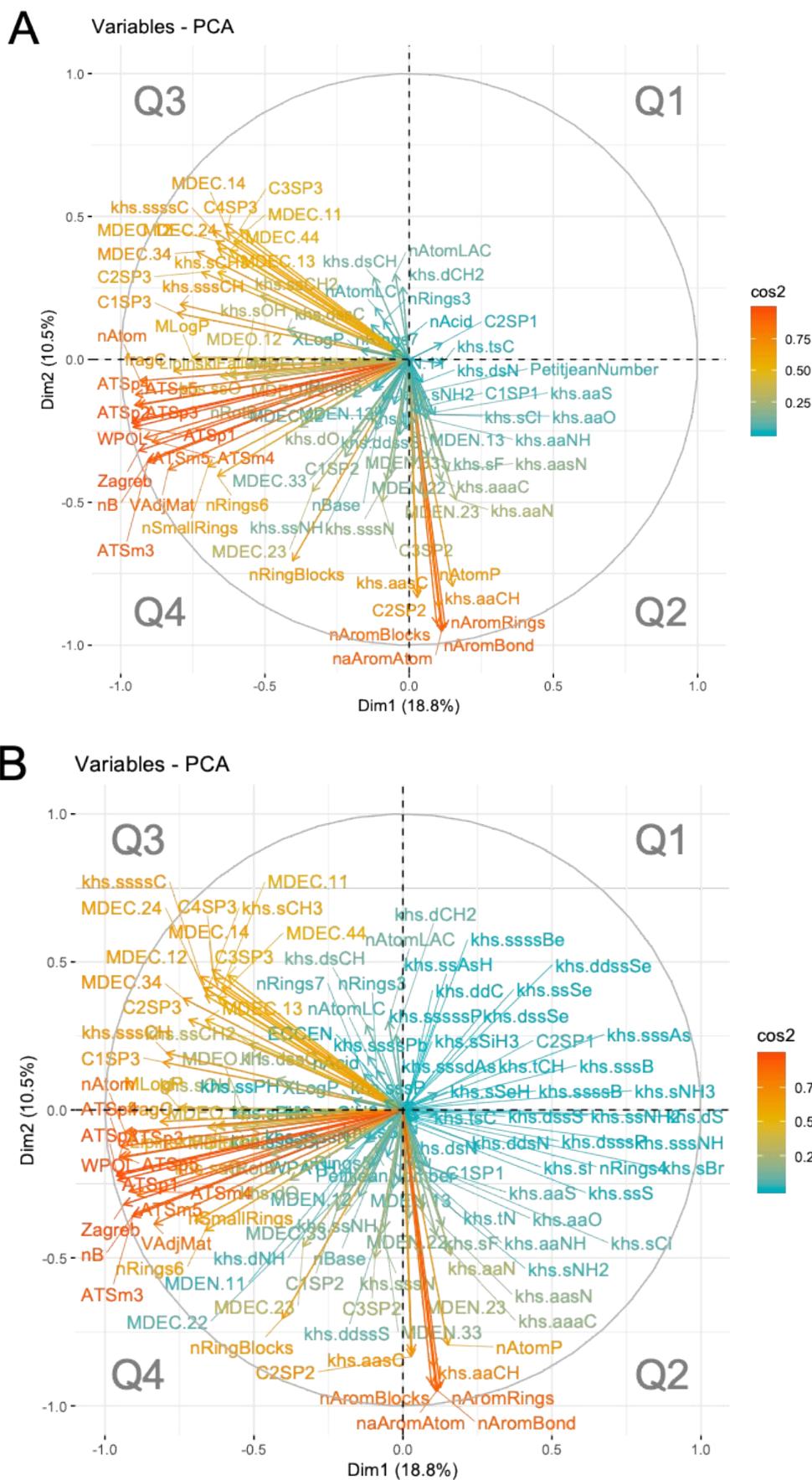
Figure S1. Principal component analysis (PCA) based on molecular descriptors of all datasets used. A) variable map, circle of correlation of the most important molecular descriptors; B) variable map, circle of correlation of all molecular descriptors. The PC1 and PC2 axes that separate the four quadrants (Q1 to Q4) are emphasized by grey dotted lines.

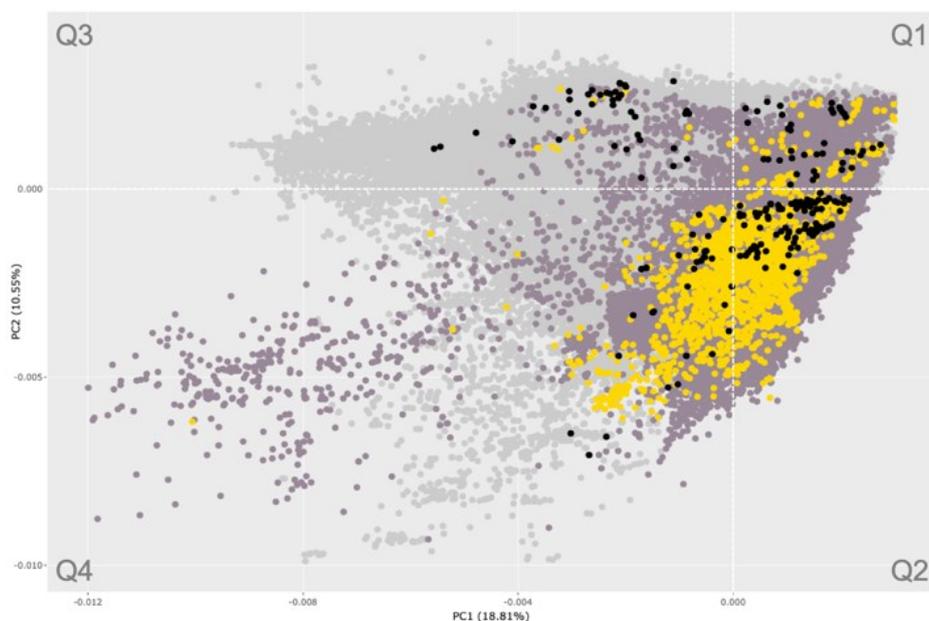## 2.2. Additional chemical space plot highlighting the ASD and iPPI-DB datasets

Figure S2. Principal component analysis (PCA) based on molecular descriptors of all datasets showing the chemical space plots UNPD_ISDB (grey dots), ASD (non-drugs, mauve dots), iPPI-DB (yellow dots) and reviewed alloNPs (black dots). The PC1 and PC2 axes that separate the four quadrants (Q1 to Q4) are emphasized by white dotted lines.

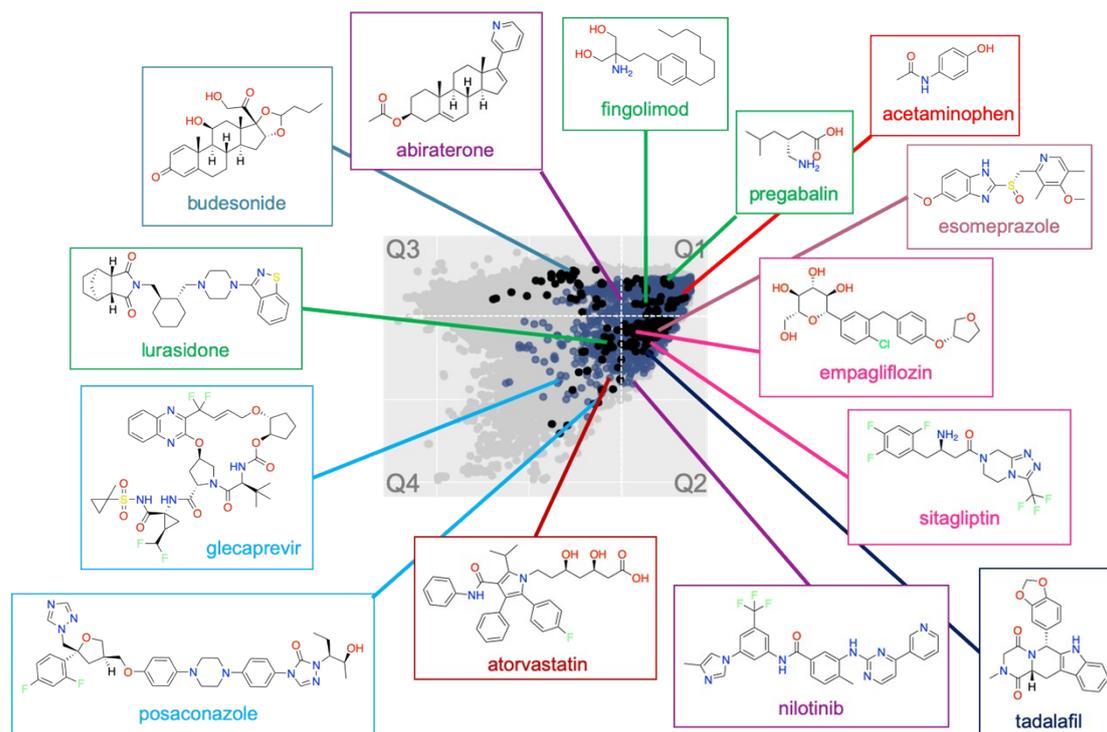## 2.3. FDA-approved drugs in the chemical space

Figure S3. Localization of selected FDA-approved drugs (among the 200 top selling small molecule drugs in 2018)[16] in the chemical space. Principal component analysis (PCA) based on molecular descriptors of all datasets showing the chemical space plots UNPD_ISDB (grey dots), FDA-approved drugs (blue dots), and reviewed alloNPs (black dots). The PC1 and PC2 axes that separate the four quadrants (Q1 to Q4) are emphasized by white dotted lines. Colours of drug names indicate therapeutic indication: oncology (violet), respiratory disorders (blue), neurological disorders (green), anti-inflammatory (bright red), gastrointestinal disorders (prune), diabetes (pink), sexual health (dark blue), cardiovascular (dark red), and infectious diseases (cyan).

---

[16] N. A. McGrath, M. Brichacek and J. T. Njardarson, *J. Chem. Educ.*, **2010**, *87*, 1348–1349. Latest data retrieved from the Njardarson group website. https://njardarson.lab.arizona.edu/content/top-pharmaceuticals-poster (accessed on January 16th 2020).

3. Groups of allosteric mechanisms found for natural allosteric modulators

Several models have been reported to mathematically explain allostery, the Monod-Wyman-Changeux (MWC model)[17] two-state model considers pre-existent conformers of proteins that are selected by ligands; the Koshland, Nemethy and Filmer (KNF model)[18] induced-fit model takes into account protein plasticity and changes promoted by ligands; and the Cuendet, Weinstein, and LeVine[19] ensemble model defines conformational landscapes that explain allostery in a free-energy basis. In recent years, when more allosteric mechanisms have been elucidated in a structural and dynamics base, it has been a trend to merge or to include more variables into these models.

Allostery is easier explained for enzymes, in which there is a well-defined catalytic site, where a substrate binds. The catalytic site generally exists in more than one local conformation: one active conformation capable of performing substrate catalysis, and at least a second conformation that is catalytically inactive. Substrate ("S" in Figure S4) binding to the enzyme traps or stabilizes the active conformation, allowing catalysis of the substrate. Competitive inhibitors ("I") also bind to the same site as the substrate, avoiding substrate binding and, therefore, catalysis, even if the enzyme's active conformation is achieved. Allosteric modulators ("*") can act following two major alterations in enzymes: i) local or ii) global. In both cases, the allosteric modulator binds to a secondary site away from the catalytic site – the allosteric site ("A") of binding. The allosteric site can occur a dozen of Angstroms away from the catalytic site.

By analysing the structural and dynamics bases of natural allosteric modulators reported over the last decade we could classify them into three major groups: type "i" (local alterations), type "ii" (global alterations) and "PPI" type. Receptors, as GPCRs, use a combination of the three phenomena. In type "i" allosteric modulation, ligand binding provokes local alterations at the catalytic site in a way that abolishes enzyme activity. The local alterations commonly involve the incorrect stabilization of protein loops or residues involved in enzyme catalysis. This is observed for example for PTP1B[20] and IsPD[21] enzymes.

On the other hand, type "ii" allosteric modulation involves global changes in protein structure and dynamics, the latter being the major component. A combination of "in solution" studies, crystallographic snapshots[22,23] of enzyme
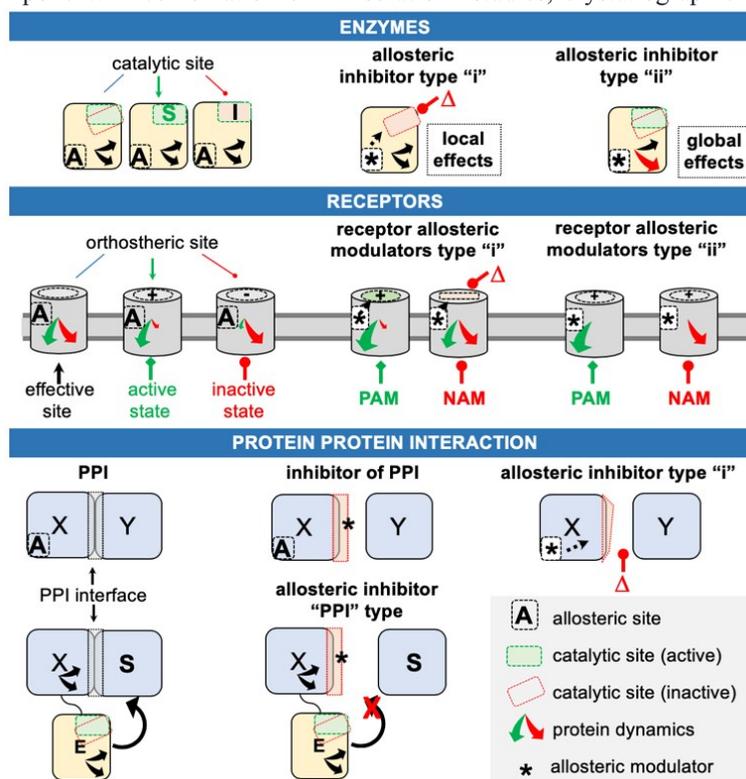

Figure S4. Types of allosteric mechanisms identified here after examining the structural and dynamics bases of natural allosteric modulators.

[19] M. A. Cuendet, H. Weinstein and M. V. LeVine, *J. Chem. Theory Comput.*, **2016**, *12*, 5758–5767.

[20] N. Krishnan, D. Koveal, D. H. Miller, B. Xue, S. D. Akshinthala, J. Kragelj, M. R. Jensen, C. M. Gauss, R. Page, M. Blackledge, S. K. Muthuswamy, W. Peti and N. K. Tonks, *Nat. Chem. Biol.*, **2014**, *10*, 558–566.

[21] A. Kunfermann, M. Witschel, B. Illarionov, R. Martin, M. Rottmann, H. W. Höffken, M. Seet, W. Eisenreich, H. J. Knölker, M. Fischer, A. Bacher, M. Groll and F. Diederich, *Angew. Chemie - Int. Ed.*, **2014**, *53*, 2235–2239.

[22] M. Arciniega, P. Beck, O. F. Lange, M. Groll and R. Huber, *Proc. Natl. Acad. Sci.*, **2014**, *111*, 9479–9484.

[23] H. P. Morgan, I. W. McNae, M. W. Nowicki, V. Hannaert, P. A. M. Michels, L. A. Fothergill-Gilmore and M. D. Walkinshaw, *J. Biol. Chem.*, **2010**, *285*, 12892–12898.

conformers, solvation and dynamics calculations might be necessary to dissect the logic of allosteric modulation type "ii".[24] Here we observed that and illustrated for the MRSA pyruvate kinase[25] and the 20S proteasome.[26]

Variations of type "i" and type "ii" allosteric modulation found in enzymes are observed in other proteins. A specific nomenclature is used for receptors, in particular for GPCRs - Figure S4 middle panel. The concept of NAM (negative allosteric modulation) and PAM (positive allosteric modulation) relates the cell-effect the allosteric modulator provokes. A very explanative review on this topic has been prepared by Changeux and Christopoulos,[27] and the reader is referred to it.

Another variation of allosteric modulation type "i" and "ii" involves protein-protein interactions (PPIs). PPIs display a central role in biology, involving major steps in intracellular signalling, transcription, and enzyme-substrate recognition, when the enzyme's substrate is another protein. In the latter case, the modulator of a PPI is also considered an allosteric modulator, once its binding site is located far away from the catalytic site. This type of allosteric modulator is classified here as allosteric modulator "PPI type". This was observed by gossypol interaction with the PARP1 PPI interface, in this case however forming BRC dimers hampering the interaction of the BRC with PARP1 substrates.[28] Further, the change in the PPI interface can be indirect, being this change provoked by the allosteric modulator binding to the allosteric site, far away from the PPI interface. This can be viewed as a type "i" allosteric modulation, where the orthosteric site, however, is the PPI and not the catalytic site - Figure S4 bottom panel. The latter was observed in the CBP coactivator case exemplified in the main text.

[24] V. J. Hilser, J. O. Wrabl and H. N. Motlagh, *Annu. Rev. Biophys.*, **2012**, *41*, 585–609.

[25] R. Zoraghi, L. Worrall, R. H. See, W. Strangman, W. L. Popplewell, H. Gong, T. Samaai, R. D. Swayze, S. Kaur, M. Vuckovic, B. B. Finlay, R. C. Brunham, W. R. McMaster, M. T. Davies-Coleman, N. C. Strynadka, R. J. Andersen and N. E. Reiner, *J. Biol. Chem.*, **2011**, *286*, 44716–44725.

[26] P. A. Osmulski and M. Gaczynska, *Mol. Pharmacol.*, **2013**, *84*, 104–113.

[27] J. P. Changeux and A. Christopoulos, *Diabetes, Obes. Metab.*, **2017**, *19*, 4–21.

[28] Na Z, Peng B, Ng S, Pan S, Lee JS, Shen HM, Yao SQ. *Angew Chem Int Ed Engl.*, **2015,** *54(8)*, 2515-2519.