THE JOURNAL OF PHYSICAL CHEMISTRY LETTERS
A JOURNAL OF THE AMERICAN CHEMICAL SOCIETY

pubs.acs.org/JPCL

Letter

Machine Learning Prediction of Allosteric Drug Activity from Molecular Dynamics

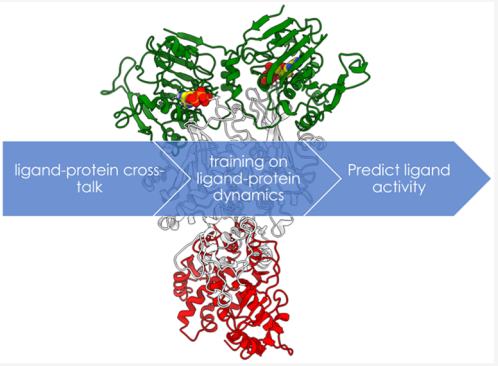
Filippo Marchetti, Elisabetta Moroni, Alessandro Pandini,* and Giorgio Colombo*

Cite This: *J. Phys. Chem. Lett.* 2021, 12, 3724–3732

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Allosteric drugs have been attracting increasing interest over the past few years. In this context, it is common practice to use high-throughput screening for the discovery of non-natural allosteric drugs. While the discovery stage is supported by a growing amount of biological information and increasing computing power, major challenges still remain in selecting allosteric ligands and predicting their effect on the target protein's function. Indeed, allosteric compounds can act both as inhibitors and activators of biological responses. Computational approaches to the problem have focused on variations on the theme of molecular docking coupled to molecular dynamics with the aim of recovering information on the (long-range) modulation typical of allosteric proteins.



Here, we present a protocol that combines docking-based screening, information on the conformational dynamics of the protein, and machine learning (ML) to classify ligands of the molecular chaperon Hsp90 as activators or inhibitors. To this end, we develop a classifier of activation/inhibition of Hsp90 allosteric ligands that is trained on data from a panel of ensemble docking results. The data set for this study is built from a database of 133 known Hsp90 ligands.

Three different ML methods are compared with the best-performing algorithm, achieving an average balanced accuracy of 0.90 (over 10-fold cross-validation) in correctly separating inhibitors from activators. A comparison with a direct classification of the chemical properties of ligands suggests that the ML prediction is not dependent on the similarity among the molecular structures but recovers hidden similarities in functional effects of different ligands.

The improved knowledge of gene organization coupled with the advances in gene editing and structural analysis methods can potentially start a whole new era in drug discovery.[1,2] In particular, improved target identification can shed light on biomolecules whose perturbation via small-molecule binding results in a functional response, transforming a disease phenotype into a normal one. The extraordinary complexity of biochemical networks in healthy and disease conditions[3,4] and the costs associated with drug discovery are however hampering the advent of this new era of therapeutics, as shown by the relatively low numbers of new drugs approved in the past few years.[5,6] Most drug discovery efforts aim at targeting the active sites of enzymes or the orthosteric sites of regulatory proteins. Because of the evolutionary and structural conservation of such sites across the proteome, issues related to selectivity, off-target effects, and development of drug resistance have started to appear.

In this context, allosteric ligands have recently emerged as a viable complement or alternative to active-site directed molecules, with novel potential as drug candidates or chemical tools.[7−10] Allosteric ligands bind to sites that are generally distinct and distal from the classic orthosteric ones. In doing so, they can perturb the target not only by inhibition but also through modulation or activation of specific functions. This represents an advantage in terms of fundamental and applicative perspectives. In fundamental research, chemical modulators (effectors) can be used to direct signaling pathways and whole cells toward desired functional states, representing important tools for understanding the roles of specific biomolecules in complex biochemical networks.[11,12] In biomedical applications, since they target sites that are generally less evolutionarily conserved, allosteric ligands can be highly selective, even among different members of the same protein family,[13] providing new opportunities for therapeutic discovery.

To date, most (non-natural) allosteric ligands/drugs have been discovered using high-throughput screening. The ever growing amount of sequences and structural information combined with the increases in computing power and the

Received: January 6, 2021
Accepted: April 5, 2021
Published: April 12, 2021

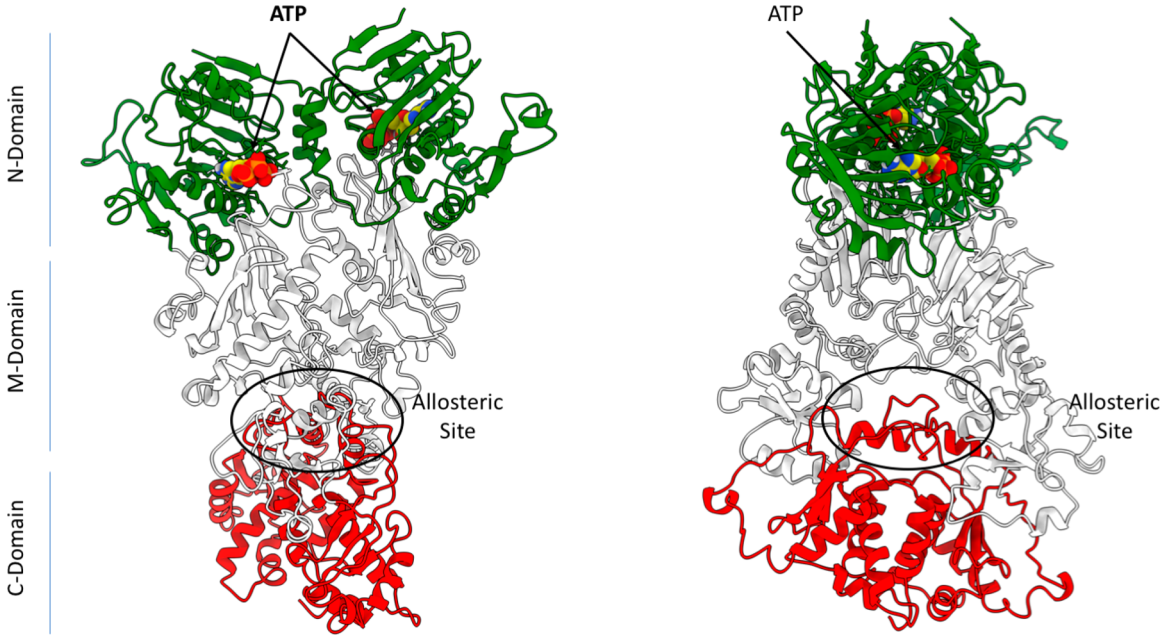


Figure 1. 3D structure and domain organization in Hsp90. The N-domain is colored in green, the middle in white ,and the C-domain in red. The ATP molecule is shown in its binding site in van der Waals representations with atom type coloring. The allosteric site is also highlighted.

improvement of predictive algorithms are starting to facilitate the discovery of allosteric modulators, but major challenges remain to develop approaches focused on rational drug design.

Computational approaches to the problem have focused on variations on the theme of molecular docking. Binding affinities predicted by docking simulations are routinely used in virtual screening to estimate relative ligand rankings and to inform further steps in lead identification.[14,15] Efficient screening of large libraries of compounds is achieved by the use of approximate scoring functions and simplified strategies for conformational sampling.[16] Typically, a static model of the target structure is used. However, recently the influence of protein dynamics on the recognition process has been more accurately modeled using ensemble strategies.[17−22] These strategies involve the docking of a molecular ligand libraries over an ensemble of selected geometries of the protein, creating a more realistic representation of the ligand bound to the different expected conformations of the target. The use of an ensemble of conformations reduces the dependence of the docking results on the target structure.[23] Ensembles can be extracted from unbiased molecular simulations of apo structures[24] and more often by sampling of protein conformations from holo structures containing first-generation ligands.[25] Under the assumption of conformational selection, a set of different ensembles representing different binding states would have selective preferential binding for different ligands. On the basis of this hypothesis, previous studies have used "a panel of ensembles" for virtual screening,[26] whereby a vector of binding affinities against the panel is used to generate a specific fingerprint for each ligand.

This type of data has high dimensionality both in the chemical and conformational space and is best suited for analysis using ML methods, which have been increasingly adopted in drug discovery studies. Indeed, they contributed to the improvement of performance in virtual screening studies[27−29] and they have been effectively used in the enhancement of structural-based virtual screening and scoring.[30,31] ML methods are mostly data-driven, and their performance is often dependent on the size and quality of the data set. To this end, they may present limited transferability, and care is required in reporting results and the scope of applicability.

The combination of ML with molecular simulations can dramatically advance the process of selection of allosteric ligands with a desired impact on the function of the target.[32] Indeed, a major limitation in docking simulations is the lack of information on the functional consequences of the allosteric binding event. While relative binding affinities and geometries can be reproduced close to experimental accuracy, there is no predictive score to discriminate inhibitors from activators, agonists from antagonists or partial agonists.[33] Experimental assays typically report on the orthosteric function, in most cases by direct measurement of a relevant biochemical parameter that involves the active/orthosteric site. This may not necessarily reflect the affinity of binding at the allosteric site.[34−36] In most cases, binding is only one aspect of an intricate interplay of structural and dynamic factors that emerges from the cross-talk between the allosteric ligand and the protein and define functional responses. As a consequence, the derivation of structure−activity relationships (SARs) for allosteric ligands is typically much more complex than for orthosteric ones.

This unmet challenge calls for new approaches that integrate information on binding, conformational dynamics, and biological activity because the desired readout of the binding event is a change of functional state in the protein that is not directly or easily modeled by single docking calculations.

Here, to progress along this fascinating avenue, we explore the potential of ML models trained on molecular simulations to predict the functional effect of allosteric ligands on proteins. Allosteric ligands can either activate or inhibit protein function. As a test case, we focus on the difficult case represented by the Hsp90 chaperone system, a molecular machinery essential for cell development and maintenance that works by facilitating the folding of a broad spectrum of clients.[37−42] Proteins of the Hsp90 family (Hsp90 in the cytosol, Grp94 in the ER and

https://doi.org/10.1021/acs.jpclett.1c00045
J. Phys. Chem. Lett. 2021, 12, 3724−3732

Trap1 in mitochondria) are homodimers with two chains consisting of three globular domains, the N-terminal (NTD), middle (M), and C-terminal (CTD). The functions of the chaperone are regulated by ATP hydrolysis in the NTD, where ATP processing is coupled to Hsp90 conformational reorganization and consequent client remodelling (see Figure 1). Early work by Neckers' group and recent computational studies reported an allosteric site at the boundary between the M- and C-terminal domains that modulates ATP-related functionalities[10,43−45] (see Figure 1). The discovery of this allosteric site facilitated the development of different series of allosteric ligands that are able to perturb Hsp90 mechanisms, by either inhibition or activation of ATP processing. Kinetic and biochemical data indicated that the functional effects of the ligands are critically coupled to their influence on the conformational dynamics of the protein.

In this work, we ask whether we can develop a reliable predictor of activation/inhibition for Hsp90 allosteric ligands. Model training is driven by ensemble-based structural, dynamic, and energetic characterization of allosteric binding.

Our approach to classify allosteric ligands as activators or inhibitors of ATP hydrolysis in Hsp90 entails three steps (see Figure 2).

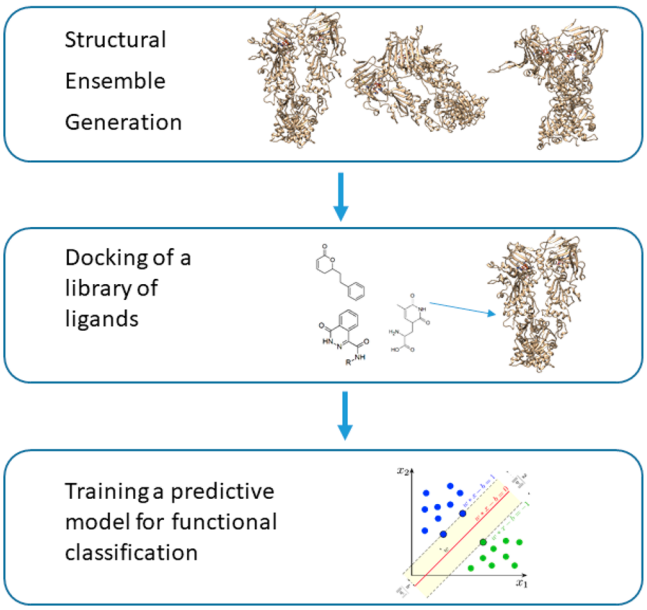


Figure 2. Simplified scheme of the MD-ML strategy. Schematic representation of the protocol followed in this work, which entails the generation of a structural ensemble, the docking of ligand libraries, and the training of learning algorithms for functional classification.

First, a panel of structural ensembles is generated by cluster analysis of conformations from molecular dynamics simulations of representative holo structures, in which Hsp90 is bound to ATP in the N-terminal domain and to an allosteric effector in the allosteric site. Then a library of allosteric compounds is docked against the Hsp90 structural panel. Finally, a predictive model for functional classification of the allosteric ligands is trained taking into account the structural, dynamic, and energetic properties of the resulting complexes. From the literature, we collected 133 compounds with known activity against Hsp90, comprising 49 inhibitors and 84 activators (see supplementary Figure S1). This data set was used to train and test the predictive model. The protein conformational ensembles for docking were generated by atomistic molecular dynamics simulations in explicit water of Hsp90 in complex with three different ligands: one activator (CC26) and two inhibitors (ND2 and Novobiocin) (see supplementary Figure S2).

We note here that in our model, the dynamics of the protein and potential allosteric effects determined by the cross-talk between the ligand and the chaperone is taken into account explicitly in these preliminary simulations with the three representative of activators and inhibitors. In this context, it is worth noting that our work aims to investigate how short time-scale changes in the structural dynamics of the chaperone dimer in the presence of small molecule effectors may determine the onset of the motions that are eventually relevant for function. The underlying hypothesis is that nanosecond time scale residue fluctuations of Hsp90 in regions that are specifically responsive to the presence of ligands may facilitate the large-scale domain rearrangements that lead to a functionally competent/incompetent state. These concepts were previously probed via computational and experimental approaches.[46]

To keep the generation of the structural ensembles independent from the data set used for training, these ligands were not included in the training and test data sets.

Each replica of molecular dynamics was run for 400 ns, saving structures every 100 ps, and the resulting trajectories were combined into a single metatrajectory. The panel of structural ensembles for docking was built to take into account the conformational variations induced by the ligands and approximate the most relevant states in Hsp90 functionally oriented dynamics. To this end and to qualitatively account for the cross-talk between the presence of a ligand and the different domains of Hsp90, geometrical cluster analysis of the metatrajectory was repeated using four different reference frameworks: the backbone atoms of N-terminal domain (Clust-N); the backbone atoms of middle domain (Clust-M); the backbone atoms of N-terminal and middle domains (Clust-NM), and the backbone atoms of middle and C-terminal domains (Clust-MC). In addition to these domain-based frameworks, a cluster analysis of the allosteric site was performed, where the ligand binding site was defined as the ensemble of residues that are within 1 nm of any bound allosteric ligand in at least 75% of all visited structures collected in the metatrajectory. This latter criterion was used to consider the most relevant local interactions between residues in and around the binding pocket and the ligand. It is worth underlining here that, during MD, the spectrum of contacts dynamically evolves.

Next, the representative structures from the three most populated clusters for each of the four domain-based ensembles were selected as a target for docking experiments. The total number of structures in the three most populated clusters always account for at least 45% of the metatrajectory. In addition, the two main representative structures resulting from the allosteric-site based clustering were added. Two structures were enough to recapitulate more than 95% of the structural variability observed in the pocket.

Cluster analysis of the molecular dynamics metatrajectory yielded a total of 14 representative protein structures for the following step of docking. This collection was generated to capture the propensity of Hsp90 to populate conformations potentially endowed with different functional properties. After docking the ligand library to each of the selected representative

https://doi.org/10.1021/acs.jpclett.1c00045
J. Phys. Chem. Lett. 2021, 12, 3724−3732

structures, three measures were calculated for every resulting complex: the docking score of the best pose for every representative structure; the root-mean-square (RMS) of the docking score for the 10 best poses; and finally, the RMSD on the atomic positions of the first 10 poses, reporting on the structural adaptation within the pocket. Ten poses per ligand were selected as a compromise to provide a tractable and easy to visualize number of configurations, while capturing positional adaptation of the molecules to a changing binding site. A total of 42 features was thus used for ML prediction.

The underlining hypothesis of our study is that features describing the docking results of a ligand against a panel of distinct conformational ensembles can be used as "dynamic fingerprints" of its functional effect on the protein. We tested this hypothesis under three assumptions: (1) the separation of activators and inhibitors cannot be directly detected in the feature space by cluster analysis; (2) the separation of activators and inhibitors requires modeling a complex relationship by supervised learning; and (3) the separation cannot be trivially obtained by use of small molecule fingerprints in the absence of information on the protein structure and dynamics.

None of the features described above can independently be used as a classifier and directly separate inhibitors from activators. This is evident from the distribution of values for every single feature against the two known ligand classes: in all the cases, the pair of per-class distributions overlap (see boxplot in Supplementary Figure 3). This suggests that a model based on the combination of these features is required to discriminate between the two classes. The first step is to test if the separation of the two groups of ligands can be directly detected with an unsupervised learning approach.

To this end, cluster analysis was performed to assess if a data segmentation compatible with the two functional classes of ligands (activators/inhibitors) can be detected. Two different algorithms were used: k-means and agglomerative hierarchical clustering. The target cluster numbers could be set to 2, but we adopted an unbiased approach and explored values between 2 and 6. The ability to correctly separate ligand classes in the clusters was estimated by cluster purity, which has values between 0 (when the class labels are completely mixed in the clusters) and 1 (clusters composed by only one class). Both algorithms have similar purity values; in particular, when 2 clusters are considered, the purity is low (0.66 for K-means and 0.69 for hierarchical), and with more clusters, the purity increases, remaining below 0.80 (for 4 clusters: k-means have 0.78 of purity and hierarchical have 0.79). The increased purity is due to the reduced size of clusters that helps adapt to the class separation. Yet, the value in the case of 2 clusters reveals that is difficult to detect a segmentation of the compounds in the functional classes directly by cluster analysis. This suggest that it is not possible to automatically partition the space of the data to identify inhibitors and activators. A model trained on properties from the different binding conformations is therefore needed.

In this framework, a classification model was built using supervised learning. The model is trained to predict class labels describing the functional effect of the ligand (i.e., activation or inhibition). Three widely used algorithms were compared: Logistic Regression (LR) as a baseline, Support Vector Machine (SVM), and Random Forest (RF). The performances of the three methods was compared after training and testing using the holdout method, where the data set is randomly split in training set and test set with the proportion of 70% and 30% respectively. The performance in prediction is reported in Table 1.

Table 1. Performance of ML Approaches: Values of Balanced Accuracy, Precision and Recall, False Positive Rate, and False Negative Rate for All Three ML Models Tested in the Paper

measure	LR	SVM	RF
balanced accuracy	0.88	0.89	0.74
precision (positive predictive value)	0.92	0.96	0.81
recall (true positive rate)	0.88	0.85	0.85
false positive rate	0.12	0.07	0.37
false negative rate	0.12	0.15	0.15

LR and SVM show similar performances while RF has poorer performance. Nevertheless, all three methods show a better classification power compared with the cluster segmentation. A 10-fold cross-validation without shuffling was performed to exclude any bias due to the simple holdout split and to further compare the methods. This approach also highlighted possible variability across the data sets and facilitated interpreting the performance with more insight on the chemical features of the molecules (see below).

SVM shows the best performance with an average balanced accuracy of 0.90, compared with 0.87 (LR) and 0.79 (RF). In Figure 3, per-fold balanced accuracy is reported. Only for one fold, values are below 0.8. The results show consistency in performance by SVM across the set. To confirm that the model is properly trained, its convergence was assessed at the increase of the training set. For each subset size ranging from 20 to 100%, 100 random samples were generated. SVM models were trained and tested with holdout. Performance was evaluated as median accuracy over the 100 random samples. The accuracy converged to 0.89 for the subset at 80%. This suggests that with a data set of ∼100 compounds the model can be built with confidence (see Supplementary Figure 4).

Finally, the possible dominance of one type of ensemble feature (docking score, rmsd, rms) in the prediction was assessed by selectively excluding each feature in turn and repeated cross-validation. In each case the variation in the average performance was not statically significant (i.e., z test score below 1; see Supplementary Figure 5), and therefore, no feature was detected as dominant.

The classification model trained on docking against the panel of representative conformations does not directly account of chemometrics properties of the ligands. In the context of compound selection, it is interesting to compare the classification model with a direct analysis of the chemical properties of the compounds. This is to assess if correct classification can be obtained by small molecule fingerprints in the absence of information on protein structure or dynamics.

Our data set comprises molecules representative of different chemotypes (Supporting Figure 1). It may be possible to qualitatively cluster these molecules with respect to shared scaffolds: in our case, this results in eight different groups (Figure 4).

Yet, the compounds can still display substantial differences in their substituents in terms of dimensions, charges, and functional groups. Therefore, a classification based only on the core of the molecules would give only a rough estimate of the chemical variability in the data set. For this reason, to explore

https://doi.org/10.1021/acs.jpclett.1c00045
J. Phys. Chem. Lett. 2021, 12, 3724−3732

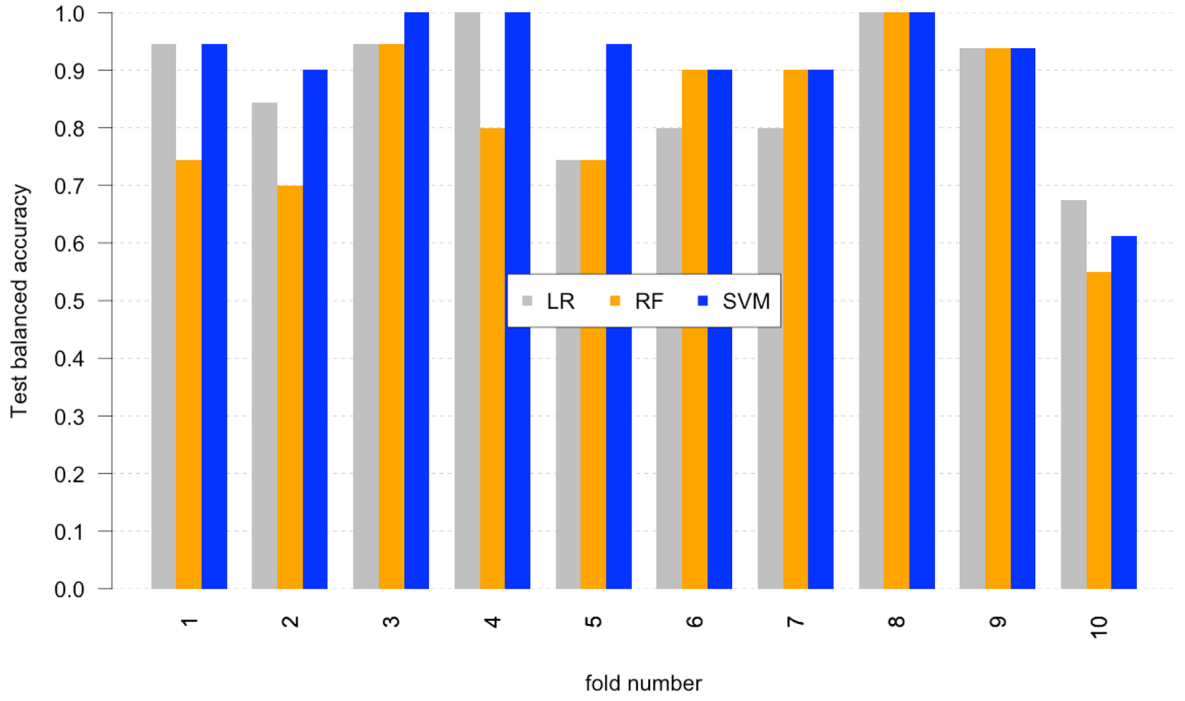


Figure 3. Performances of the 10-fold cross-validation for all the models. The values of balanced accuracy for every fold presented: the values for Logistic Regression are in gray, and Random Forest are in orange and SVM in blue.

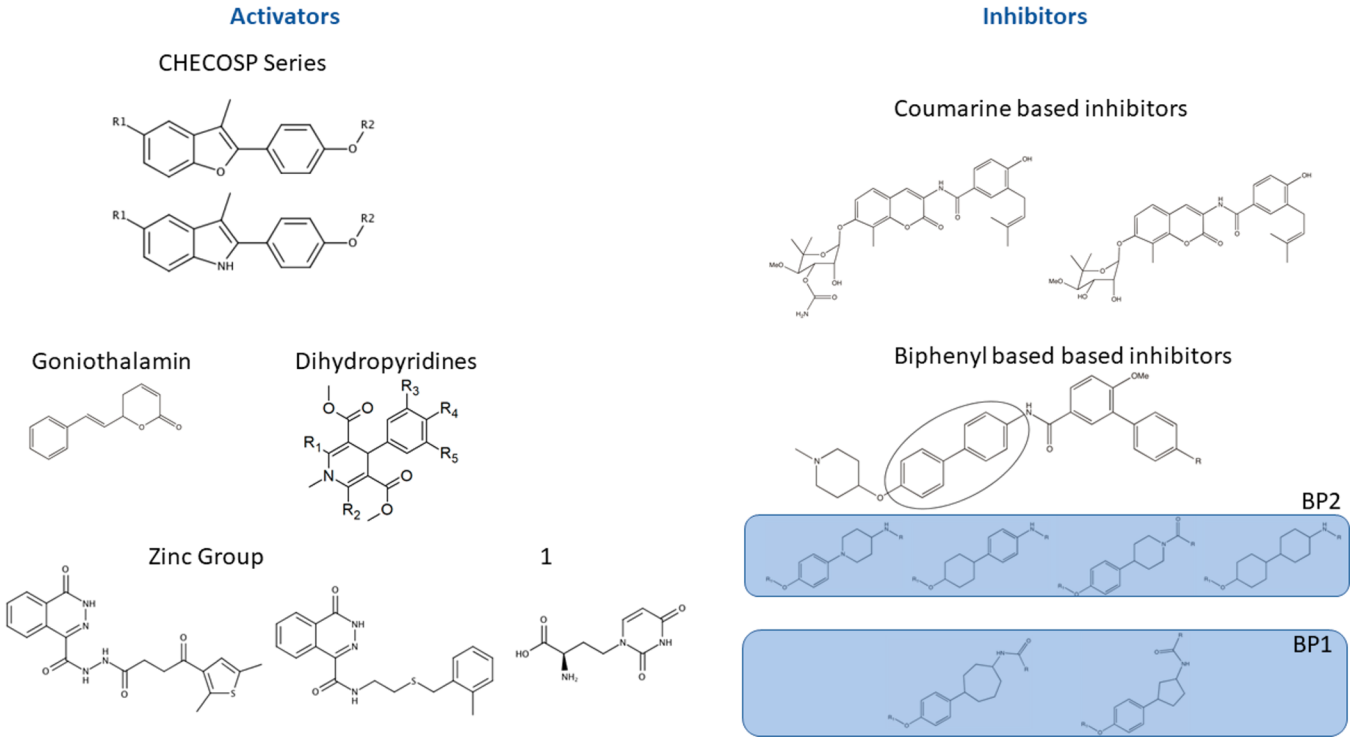


Figure 4. Subdivision of the studied molecules in distinct groups. The 2D structure of the scaffolds are divided in eight groups according to a scaffold-similarity criterion. From left to right: CheCOSP molecules (CC), coumarine-based inhibitors (CB), goniothalamin (GT), dihydropyridines (DP), the biphenyl inhibitors set is split in two groups (BP1 and BP2), the Zinc Group (Z), and last the compound labeled with 1 makes his own group (Unk).

the possibility of classifying the function of molecules based only on their chemical properties, we used a more quantitative method based on cheminformatics similarity criteria. A common method to evaluate the similarity among compounds is to compute the Tanimoto coefficient on molecular fingerprints.[47] The efficacy of similarity algorithms tends to vary with biological activity; therefore, the choice of the fingerprint model usually depends on the system under study. Here, our aim is specifically to introduce a metric for the comparison with our ML-dynamics based predictions. Since

https://doi.org/10.1021/acs.jpclett.1c00045
J. Phys. Chem. Lett. 2021, 12, 3724−3732

the best fingerprint model for our data set is not known, we tried two widely used methods: ECFP, a method that maps a molecule with a set of fragments radially grown from each heavy atom; and MACCS, which accounts for the presence/absence of specific structural features.[48] In both cases, the molecules are clustered using k-means algorithm with a cluster number varying from 2 to 6 (Table 2). The ECFP fingerprint works better in separating the compounds between activators and inhibitors when only 2 clusters are chosen, while with 3 or 4 clusters, the separation is similar. With 3 clusters, we found that the CheCOSP[20] group is separated. This consists only of activators validated by experimental characterization. The result shows that, despite a shared scaffold, there is a substantial chemical variety in the group. Segmentation for higher cluster numbers does not clearly lead to any grouping consistent with the chemical properties of the ligands.

Table 2. Performance of the Fingerprinting Methods

	MACCS		ECFP	
	activator	inhibitor	activator	inhibitor
K = 2				
cluster 1	33	48	12	49
cluster 2	51	1	72	0
K = 3				
cluster 1	45	1	67	0
cluster 2	0	47	17	2
cluster 3	39	1	0	47
K = 4				
cluster 1	16	0	42	0
cluster 2	0	47	14	2
cluster 3	36	1	0	47
cluster 4	32	1	28	0
K = 5				
cluster 1	44	0	31	0
cluster 2	1	32	0	42
cluster 3	0	29	0	28
cluster 4	0	13	16	0
cluster 5	4	10	2	14
K = 6				
cluster 1	44	0	2	14
cluster 2	1	32	0	42
cluster 3	0	29	16	0
cluster 4	0	12	31	0
cluster 5	3	6	0	14
cluster 6	1	5	0	14

In Table 2, we report the performances of MACCS and ECFP fingerprints in recognizing and assigning activators and inhibitors, obtained with k-means clustering.

The best result obtained by ECFP fingerprint on two clusters was compared with the best ML predictive model obtained by SVM (all data in Supporting Table 1). The comparison was broken down by chemical groups to explore how the two approaches perform on different subclasses of ligands. In Figure 5, we report the fraction of correct classifications for every group in our data set. For the group of three inhibitors (BP1, BP2, and CB), a high fraction of correctly classified is observed for ECFP, meaning that inhibitors have good chemical similarity, whereas for activators, the fraction for ECFP is high only for CC group. In all the other groups (Z, DP, Unk, and GT), the fraction is 0. Interestingly, the SVM model correctly predicts as activators even the groups with low similarity with CC (the group most extensively characterized at the experimental level). In this context, we notice that SVM still correctly predicts group Z to 0.3 (0.0 in the case of fingerprints), DP to 0.6, Unk and GT to 1. In contrast, inhibitors of the CB groups have good similarity with the rest of the inhibitors, but they are not correctly predicted by SVM.

Overall, the results of this comparative analysis suggest that the characterization of allosteric binding with the partner protein, which reverberates the cross-talk between the ligand and the receptor, captures the main structural and dynamic determinants at the basis of allosteric modulation. On the one

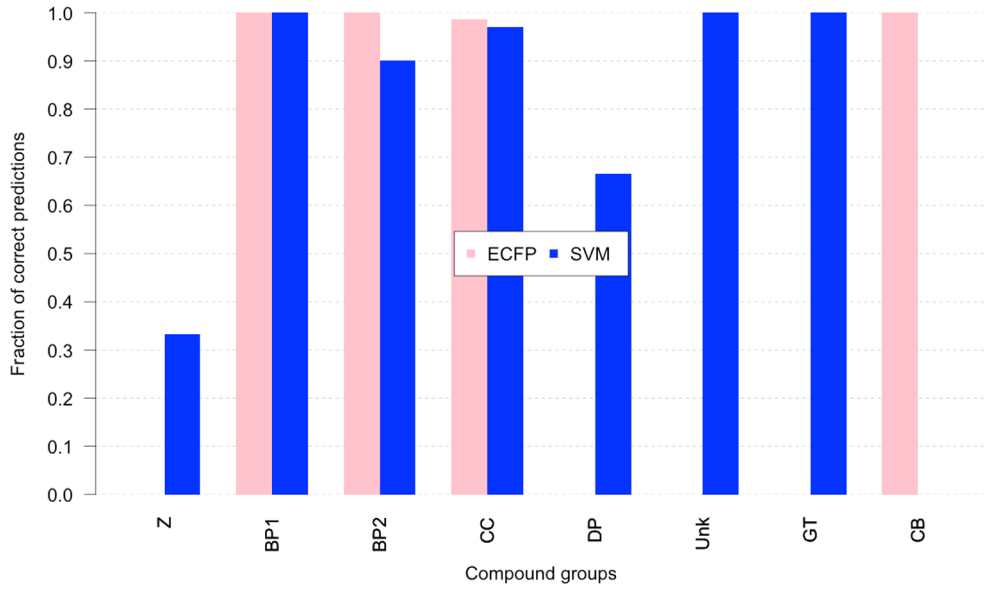


Figure 5. Performance comparison between SVM and ECFP. The graph reports the fraction of correct predictions obtained with the SVM method compared with the cluster separation of ECFP values. An entry of the ECFP cluster is considered correctly separated if it is located in the cluster that contains the majority of its class. The values are in pink for ECFP and in blue for SVM, the fractions are evaluated separately for every scaffold group

https://doi.org/10.1021/acs.jpclett.1c00045
J. Phys. Chem. Lett. 2021, 12, 3724−3732

hand, this approach is not dependent on the similarities among the molecular structures of the libraries of compounds under exam. On the other hand, considering that specific functionalities may determine recognition, binding, and the successive functional regulation, it is important to underline that the relevance of specific chemotypes for functional modulation emerges from the ML analysis. This aspect is aptly captured by the suitable combination of docking and Molecular Dynamics.

The most successful predictor is a learned supervised model built on features describing the protein−ligand interaction across the whole set of representative structures from the conformational panel. Attempts to use only some features or some structures leads to poorer performance. This is consistent with the current understanding that functional activation by allosteric ligands is often mediated by the ligand "selecting" some of the conformational states. Information on both selected and nonselected states is required to identify effective binding. This also suggests that the model has learned the relationship between selective binding patterns and functional effect. Therefore, the need for more sophisticated unsupervised algorithm is explained: this relationship is multivariate, not known in its analytical form and complex.

We propose that the ML strategy we have presented here, while demonstrated on a specific but highly challenging case, is not system-specific and could be extended to the study of other allosterically regulated systems: in this context, we propose our method as a valid complement to the selection of allosteric leads for potential drug-development.

MOLECULAR DYNAMICS SIMULATION AND ANALYSIS

The protein structure coordinates (PDB ID: 2CG9) for yeast Hsp90 were downloaded from the Protein Databank. Initial poses for ligand docking were derived from previously published models.[20,21,49,50] MD simulations were run with Gromacs 2018.2[51] with the Amber03 force field.[52] The protein−ligand complex was solvated with TIP3P water model in a dodecahedral box with minimal distance from the solute of 1.4 nm, and counterions were added to neutralize the system. After a minimization, the molecules were equilibrated for 100 ps in the NVT ensemble and successively in the NPT ensemble for 100 ps. The simulations were conducted at a constant temperature of 300 K and at a constant pressure of 1 bar, with a coupling time of 2 ps. The electrostatic term was described by using the particle mesh Ewald algorithm,[53] and the LINCS algorithm[54−56] was used to constrain all bond lengths. Available ATP parameters for the Amber force field[57] were used, and ligands topologies were generated using AnteChamber from the AmberTools module of AMBER18[52] suite. The atomic point charges were generated with the AM1-BCC charge model, and bonded and nonbonded parameters were automatically assigned with the combination rules defined by the AnteChamber module of the Amber Suite. For each ligand-protein complex, a 400 ns of simulation was run. Cluster analysis was performed on a combined metatrajectory of all simulations with the representative ligands. Rigid roto-translation fitting and RMSD calculations were made on α carbon atoms of secondary structure segments extracted with VMD software.[58] Clustering was performed with Gromos algorithm[59] using a cutoff between 2 and 2.5 Å.

MOLECULAR DOCKING AND FINGERPRINT ANALYSIS

All systems were prepared using the Maestro Software Suite from Schrodinger (www.schrodinger.com): Bond orders and atomic charges were assigned, and the hydrogens were added. Protonation states were evaluated on acid and basic enzymes, and hydrogen bonds were optimized. The protein was then minimized with a Cutoff of 0.3 with respect to starting configuration. The Glide[60] software was used for molecular docking: the putative binding site was mapped on a grid with dimensions of 48 A, enclosing box, and 28 A, inner box. Calculations with a fixed receptor and flexible ligand were made with standard precision (SP) modality with OPLS3e Force Field. No additional changes to default settings were made. Fingerprint similarities were computed with the Canvas program of the Schrodinger Suite, and the typing scheme is atom distinguished by functional type with no scaling in 32 bit.

SUPERVISED AND UNSUPERVISED LEARNING

In-house scripts for cluster analysis and supervised learning prediction were developed in Python using scikit-learn functions.[61] Source code is released under GNU General Public License and available at https://github.com/alepandini/LIGXF. The Logistic Regression models were trained using default settings in scikit-learn. These included L2 norm for penalty estimation with 1e-4 tolerance for stopping criteria and Limited-memory Broyden−Fletcher−Goldfarb−Shanno algorithm (LM-BFGS) for optimization. The SVM models were trained using a linear kernel and all other settings were set to default values. The RF models were trained with an increased number of trees (1000) compared with default, and the best split at decision points was selected by minimization of Gini impurity.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jpclett.1c00045.

Supplementary Table 1; Supplementary Figures S1−S5 (PDF)

AUTHOR INFORMATION

Corresponding Authors

Giorgio Colombo − *Department of Chemistry, Università Degli Studi di Pavia, 27100 Pavia, Italy; Istituto di Scienze e Tecnologie Chimiche, 20131 Milano, Italy;* orcid.org/0000-0002-1318-668X; Email: g.colombo@unipv.it

Alessandro Pandini − *Brunel University London, Uxbridge UB8 3PH, U.K.;* orcid.org/0000-0002-4158-233X; Email: alessandro.pandini@brunel.ac.uk

Authors

Filippo Marchetti − *Department of Chemistry, Università Degli Studi di Pavia, 27100 Pavia, Italy; Università Degli Studi di Milano, I-20133 Milan, Italy*

Elisabetta Moroni − *Istituto di Scienze e Tecnologie Chimiche, 20131 Milano, Italy*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jpclett.1c00045

Notes

The authors declare no competing financial interest.

https://doi.org/10.1021/acs.jpclett.1c00045
J. Phys. Chem. Lett. 2021, 12, 3724−3732

■ ACKNOWLEDGMENTS

The research leading to these results has received funding from AIRC under IG 2017 - ID. 20019 project – P.I. Colombo Giorgio; Filippo Marchetti was supported by an AIRC Fellowship. The work has been performed under the Project HPC-EUROPA3 (INFRAIA-2016-1-730897), with the support of the EC Research Innovation Action under the H2020 Programme; in particular, Filippo Marchetti gratefully acknowledges the technical support and the computer resources provided by EPCC.

■ REFERENCES

(1) van der Greef, J.; McBurney, R. N. Innovation - Rescuing drug discovery: in vivo systems pathology and systems pharmacology. *Nat. Rev. Drug Discovery* **2005**, *4* (12), 961–967.
(2) Workman, P.; Antolin, A. A.; Al-Lazikani, B. Transforming cancer drug discovery with Big Data and AI. *Expert Opin. Drug Discovery* **2019**, *14* (11), 1089–1095.
(3) Overington, J. P.; Al-Lazikani, B.; Hopkins, A. L. How many drug targets are there? *Nat. Rev. Drug Discovery* **2006**, *5* (12), 993–996.
(4) Hopkins, A. L. Network pharmacology: the next paradigm in drug discovery. *Nat. Chem. Biol.* **2008**, *4* (11), 682–690.
(5) Reidenbach, A. G.; Mesleh, M. F.; Casalena, D.; Vallabh, S. M.; Dahlin, J. L.; Leed, A. J.; Chan, A. I.; Usanov, D. L.; Yehl, J. B.; Lemke, C. T.; Campbell, A. J.; Shah, R. N.; Shrestha, O. K.; Sacher, J. R.; Rangel, V. L.; Moroco, J. A.; Sathappa, M.; Nonato, M. C.; Nguyen, K. T.; Wright, S. K.; Liu, D. R.; Wagner, F. F.; Kaushik, V. K.; Auld, D. S.; Schreiber, S. L.; Minikel, E. V. Multimodal small-molecule screening for human prion protein binders. *bioRxiv*, June 20, **2020**, DOI: 10.1101/2020.06.18.159418.
(6) Schneider, P.; Walters, W. P.; Plowright, A. T.; Sieroka, N.; Listgarten, J.; Goodnow, R. A.; Fisher, J.; Jansen, J. M.; Duca, J. S.; Rush, T. S.; Zentgraf, M.; Hill, J. E.; Krutoholow, E.; Kohler, M.; Blaney, J.; Funatsu, K.; Luebkemann, C.; Schneider, G. Rethinking drug design in the artificial intelligence era. *Nat. Rev. Drug Discovery* **2020**, *19* (5), 353–364.
(7) Wodak, S. J.; Paci, E.; Dokholyan, N. V.; Berezovsky, I. N.; Horovitz, A.; Li, J.; Hilser, V. J.; Bahar, I.; Karanicolas, J.; Stock, G.; Hamm, P.; Stote, R. H.; Eberhardt, J.; Chebaro, Y.; Dejaegere, A.; Cecchini, M.; Changeux, J.-P.; Bolhuis, P. G.; Vreede, J.; Faccioli, P.; Orioli, S.; Ravasio, R.; Yan, L.; Brito, C.; Wyart, M.; Gkeka, P.; Rivalta, I.; Palermo, G.; McCammon, J. A.; Panecka-Hofman, J.; Wade, R. C.; Di Pizio, A.; Niv, M. Y.; Nussinov, R.; Tsai, C.-J.; Jang, H.; Padhorny, D.; Kozakov, D.; McLeish, T. Allostery in Its Many Disguises: From Theory to Applications. *Structure (Oxford, U. K.)* **2019**, *27* (4), 566–578.
(8) Serapian, S. A.; Colombo, G. Designing Molecular Spanners to Throw in the Protein Networks. *Chem. - Eur. J.* **2020**, *26* (21), 4656–4670.
(9) Daura, X. Advances in the Computational Identification of Allosteric Sites and Pathways in Proteins. In *Protein Allostery in Drug Discovery*; Zhang, J., Nussinov, R., Eds.; Springer Nature: 2019.
(10) Ferraro, M.; D'Annessa, I.; Moroni, E.; Morra, G.; Paladino, A.; Rinaldi, S.; Compostella, F.; Colombo, G. Allosteric modulators of Hsp90 and Hsp70: Dynamics meets Function through Structure-Based Drug Design. *J. Med. Chem.* **2019**, *62* (1), 60–87.
(11) Zorn, J. A.; Wells, J. A. Turning enzymes ON with small molecules. *Nat. Chem. Biol.* **2010**, *6* (3), 179–188.
(12) Szilagyi, A.; Nussinov, R.; Csermely, P. Allo-Network Drugs: Extension of the Allosteric Drug Concept to Protein- Protein Interaction and Signaling Networks. *Curr. Top. Med. Chem.* **2013**, *13*, 64.
(13) Sanchez-Martin, C.; Moroni, E.; Ferraro, M.; Laquatra, C.; Cannino, G.; Masgras, I.; Negro, A.; Quadrelli, P.; Rasola, A.; Colombo, G. Rational Design of Allosteric and Selective Inhibitors of the Molecular Chaperone TRAP1. *Cell Rep.* **2020**, *31* (3), 107531.
(14) Lionta, E.; Spyrou, G.; Vassilatis, D. K.; Cournia, Z. Structure-based virtual screening for drug discovery: principles, applications and recent advances. *Curr. Top. Med. Chem.* **2014**, *14* (16), 1923–1938.
(15) Benod, C.; Carlsson, J.; Uthayaruban, R.; Hwang, P.; Irwin, J. J.; Doak, A. K.; Shoichet, B. K.; Sablin, E. P.; Fletterick, R. J. Structure-based discovery of antagonists of nuclear receptor LRH-1. *J. Biol. Chem.* **2013**, *288* (27), 19830–19844.
(16) Lyu, J.; Wang, S.; Balius, T. E.; Singh, I.; Levit, A.; Moroz, Y. S.; O'Meara, M. J.; Che, T.; Algaa, E.; Tolmachova, K.; Tolmachev, A. A.; Shoichet, B. K.; Roth, B. L.; Irwin, J. J. Ultra-large library docking for discovering new chemotypes. *Nature* **2019**, *566* (7743), 224–229.
(17) Bowman, A. L.; Nikolovska-Coleska, Z.; Zhong, H. Z.; Wang, S. M.; Carlson, H. A. Small molecule inhibitors of the MDM2-p53 interaction discovered by ensemble-based receptor models. *J. Am. Chem. Soc.* **2007**, *129* (42), 12809–12814.
(18) Nussinov, R.; Tsai, C.-J.; Jang, H. Protein ensembles link genotype to phenotype. *PLoS Comput. Biol.* **2019**, *15* (6), e1006648–e1006648.
(19) D'Annessa, I.; Sattin, S.; Tao, J. H.; Pennati, M.; Sanchez-Martin, C.; Moroni, E.; Rasola, A.; Zaffaroni, N.; Agard, D. A.; Bernardi, A.; Colombo, G. Design of Allosteric Stimulators of the Hsp90 ATPase as New Anticancer Leads. *Chem. - Eur. J.* **2017**, *23* (22), 5188–5192.
(20) Sattin, S.; Tao, J.; Vettoretti, G.; Moroni, E.; Pennati, M.; Lopergolo, A.; Morelli, L.; Bugatti, A.; Zuehlke, A.; Moses, M.; Prince, T.; Kijima, T.; Beebe, K.; Rusnati, M.; Neckers, L.; Zaffaroni, N.; Agard, D.; Bernardi, A.; Colombo, G. Activation of Hsp90 Enzymatic Activity and Conformational Dynamics through Rationally Designed Allosteric Ligands. *Chem. - Eur. J.* **2015**, *21* (39), 13598–13608.
(21) Vettoretti, G.; Moroni, E.; Sattin, S.; Tao, J.; Agard, D.; Bernardi, A.; Colombo, G. Molecular Dynamics Simulations Reveal the Mechanisms of Allosteric Activation of Hsp90 by Designed Ligands. *Sci. Rep.* **2016**, *6*, 23830.
(22) D'Annessa, I.; Raniolo, S.; Limongelli, V.; Di Marino, D.; Colombo, G. Ligand Binding, Unbinding, and Allosteric Effects: Deciphering Small-Molecule Modulation of HSP90. *J. Chem. Theory Comput.* **2019**, *15* (11), 6368–6381.
(23) Carlson, H. A.; McCammon, J. A. Accommodating protein flexibility in computational drug design. *Mol. Pharmacol.* **2000**, *57* (2), 213–218.
(24) Amaro, R. E.; Baudry, J.; Chodera, J.; Demir, Ö.; McCammon, J. A.; Miao, Y.; Smith, J. C. Ensemble Docking in Drug Discovery. *Biophys. J.* **2018**, *114* (10), 2271–2278.
(25) Cheng, L. S.; Amaro, R. E.; Xu, D.; Li, W. W.; Arzberger, P. W.; McCammon, J. A. Ensemble-Based Virtual Screening Reveals Potential Novel Antiviral Compounds for Avian Influenza Neuraminidase. *J. Med. Chem.* **2008**, *51* (13), 3878–3894.
(26) Tian, S.; Sun, H.; Pan, P.; Li, D.; Zhen, X.; Li, Y.; Hou, T. Assessing an Ensemble Docking-Based Virtual Screening Strategy for Kinase Targets by Considering Protein Flexibility. *J. Chem. Inf. Model.* **2014**, *54* (10), 2664–2679.
(27) Koutsoukas, A.; Monaghan, K. J.; Li, X.; Huan, J. Deep-learning: investigating deep neural networks hyper-parameters and comparison of performance to shallow methods for modeling bioactivity data. *J. Cheminf.* **2017**, *9* (1), 42.
(28) Baskin, I. I.; Winkler, D.; Tetko, I. V. A renaissance of neural networks in drug discovery. *Expert Opin. Drug Discovery* **2016**, *11* (8), 785–795.
(29) Mayr, A.; Klambauer, G.; Unterthiner, T.; Hochreiter, S. DeepTox: Toxicity Prediction using Deep Learning. *Front. Environ. Sci.* **2016**, *3*, 80.
(30) Morrone, J. A.; Weber, J. K.; Huynh, T.; Luo, H.; Cornell, W. D. Combining Docking Pose Rank and Structure with Deep Learning Improves Protein–Ligand Binding Mode Prediction over a Baseline Docking Approach. *J. Chem. Inf. Model.* **2020**, *60*, 4170.
(31) Wang, B.; Yan, C.; Lou, S.; Emani, P.; Li, B.; Xu, M.; Kong, X.; Meyerson, W.; Yang, Y. T.; Lee, D.; Gerstein, M. Building a Hybrid Physical-Statistical Classifier for Predicting the Effect of Variants

https://doi.org/10.1021/acs.jpclett.1c00045
J. Phys. Chem. Lett. 2021, 12, 3724–3732

Related to Protein-Drug Interactions. *Structure* **2019**, *27* (9), 1469−1481.e3.
(32) Ferraro, M.; Moroni, E.; Ippoliti, E.; Rinaldi, S.; Sanchez-Martin, C.; Rasola, A.; Pavarino, L. F.; Colombo, G. Machine Learning of Allosteric Effects: The Analysis of Ligand-Induced Dynamics to Predict Functional Effects in TRAP1. *J. Phys. Chem. B* **2021**, *125*, 101.
(33) Wagner, J. R.; Lee, C. T.; Durrant, J. D.; Malmstrom, R. D.; Feher, V. A.; Amaro, R. E. Emerging Computational Methods for the Rational Discovery of Allosteric Drugs. *Chem. Rev.* **2016**, *116* (11), 6370−6390.
(34) Pricer, R.; Gestwicki, J. E.; Mapp, A. K. From Fuzzy to Function: The New Frontier of Protein-Protein Interactions. *Acc. Chem. Res.* **2017**, *50* (3), 584−589.
(35) Rinaldi, S.; Assimon, V. A.; Young, Z. T.; Morra, G.; Shao, H.; Taylor, I. R.; Gestwicki, J. E.; Colombo, G. A Local Allosteric Network in Heat Shock Protein 70 (Hsp70) Links Inhibitor Binding to Enzyme Activity and Distal Protein-Protein Interactions. *ACS Chem. Biol.* **2018**, *13* (11), 3142−3152.
(36) Taylor, I. R.; Assimon, V. A.; Kuo, S. Y.; Rinaldi, S.; Li, X.; Young, Z. T.; Morra, G.; Green, K.; Nguyen, D.; Shao, H.; Garneau-Tsodikova, S.; Colombo, G.; Gestwicki, J. E. Tryptophan scanning mutagenesis as a way to mimic the compound-bound state and probe the selectivity of allosteric inhibitors in cells. *Chemical Science* **2020**, *11* (7), 1892−1904.
(37) Flynn, J. M.; Mishra, P.; Bolon, D. N. A. Mechanistic asymmetry in Hsp90 dimers. *J. Mol. Biol.* **2015**, *427*, 2904.
(38) Shrestha, L.; Patel, H. J.; Chiosis, G. Chemical Tools to Investigate Mechanisms Associated with HSP90 and HSP70 in Disease. *Cell Chem. Biol.* **2016**, *23* (1), 158−172.
(39) Rasola, A. HSP90 proteins in the scenario of tumor complexity. *Oncotarget* **2017**, *8* (13), 20521−20522.
(40) Neckers, L.; Blagg, B.; Haystead, T.; Trepel, J. B.; Whitesell, L.; Picard, D. Methods to validate Hsp90 inhibitor specificity, to identify off-target effects, and to rethink approaches for further clinical development. *Cell Stress Chaperones* **2018**, *23*, 467.
(41) Paladino, A.; Woodford, M. R.; Backe, S. J.; Sager, R. A.; Kancherla, P.; Daneshvar, M. A.; Chen, V. Z.; Ahanin, E. F.; Bourboulia, D.; Prodromou, C.; Bergamaschi, G.; Strada, A.; Cretich, M.; Gori, A.; Veronesi, M.; Bandiera, T.; Vanna, R.; Bratslavsky, G.; Serapian, S. A.; Mollapour, M.; Colombo, G. Chemical Perturbation of Oncogenic Protein Folding: from the Prediction of Locally Unstable Structures to the Design of Disruptors of Hsp90-Client Interactions. *Chem. - Eur. J.* **2020**, *26*, 9459.
(42) Schopf, F. H.; Biebl, M. M.; Buchner, J. The HSP90 chaperone machinery. *Nat. Rev. Mol. Cell Biol.* **2017**, *18* (6), 345−360.
(43) Marcu, M. G.; Schulte, T. W.; Neckers, L. Novobiocin and related coumarins and depletion of heat shock protein 90-dependent signaling proteins. *JNCI* **2000**, *92* (3), 242−248.
(44) Burlison, J. A.; Neckers, L.; Smith, A. B.; Maxwell, A.; Blagg, B. S. J. Novobiocin: Redesigning a DNA Gyrase Inhibitor for Selective Inhibition of Hsp90. *J. Am. Chem. Soc.* **2006**, *128* (48), 15529−15536.
(45) Sanchez-Martin, C.; Serapian, S. A.; Colombo, G.; Rasola, A. Dynamically Shaping Chaperones. Allosteric Modulators of HSP90 Family as Regulatory Tools of Cell Metabolism in Neoplastic Progression. *Front. Oncol.* **2020**, *10*, 1177.
(46) Rehn, A.; Moroni, E.; Zierer, B. K.; Tippel, F.; Morra, G.; John, C.; Richter, K.; Colombo, G.; Buchner, J. Allosteric Regulation Points Control the Conformational Dynamics of the Molecular Chaperone Hsp90. *J. Mol. Biol.* **2016**, *428* (22), 4559−4571.
(47) Bajusz, D.; Rácz, A.; Héberger, K. Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations? *J. Cheminf.* **2015**, *7* (1), 20.
(48) Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints. *J. Chem. Inf. Model.* **2010**, *50* (5), 742−754.
(49) Morra, G.; Neves, M. A. C.; Plescia, C. J.; Tsustsumi, S.; Neckers, L.; Verkhivker, G.; Altieri, D. C.; Colombo, G. Dynamics-Based Discovery of Allosteric Inhibitors: Selection of New Ligands for the C-terminal Domain of Hsp90. *J. Chem. Theory Comput.* **2010**, *6* (9), 2978−2989.
(50) Moroni, E.; Zhao, H.; Blagg, B. S.; Colombo, G. Exploiting Conformational Dynamics in Drug Discovery: Design of C-Terminal Inhibitors of Hsp90 with Improved Activities. *J. Chem. Inf. Model.* **2014**, *54*, 195.
(51) Abraham, M. J.; van der Spoel, D.; Lindahl, E.; Hess, B. GROMACS User Manual version 2019. http://www.gromacs.org (accessed 2019).
(52) Case, D. A.; Cerutti, D. S.; Cheatham, T. E. I.; Darden, T. A.; Duke, R. E.; Giese, T. J.; Gohlke, H.; Goetz, A. W.; Greene, D.; Homeyer, N.; Izadi, S.; Kovalenko, A.; Lee, T. S.; LeGrand, S.; Li, P. L. C.; Liu, J.; Luchko, T.; Luo, R.; Mermelstein, D.; Merz, K. M.; Monard, G.; Nguyen, H.; Omelyan, I.; Onufriev, A.; Pan, F.; Qi, R.; Roe, D. R.; Roitberg, A.; Sagui, C.; Simmerling, C. L.; Botello-Smith, W. M.; Swails, J.; Walker, R. C.; Wang, J.; Wolf, R. M.; Wu, X.; Xiao, L.; York, D. M.; Kollman, P. A. *AMBER 2018*; University of California: San Francisco, 2018.
(53) Darden, T.; York, D.; Pedersen, L. Particle mesh Ewald: An N-log(N) method for Ewald sums in large systems. *J. Chem. Phys.* **1993**, *98*, 10089−10092.
(54) Hess, B.; Bekker, H.; Fraaije, J. G. E. M.; Berendsen, H. J. C. A linear constraint solver for molecular simulations. *J. Comput. Chem.* **1997**, *18*, 1463−1472.
(55) Feenstra, K. A.; Hess, B.; Berendsen, H. J. C. Improving Efficiency of Large Time-scale Molecular Dynamics Simulations of Hydrogen-rich Systems. *J. Comput. Chem.* **1999**, *20*, 786−798.
(56) Hess, B. P-LINCS: A Parallel Linear Constraint Solver for Molecular Simulation. *J. Chem. Theory Comput.* **2008**, *4* (1), 116−122.
(57) Meagher, K. L.; Redman, L. T.; Carlson, H. A. Development of polyphosphate parameters for use with the AMBER force field. *J. Comput. Chem.* **2003**, *24* (9), 1016−1025.
(58) Humphrey, W.; Dalke, A.; Schulten, K. VMD: Visual molecular dynamics. *J. Mol. Graphics* **1996**, *14*, 33−38.
(59) Daura, X.; Gademann, K.; Jaun, B.; Seebach, D.; van Gunsteren, W. F.; Mark, A. E. Peptide folding: when simulation meets experiment. *Angew. Chem., Int. Ed.* **1999**, *38*, 236−240.
(60) Friesner, R. A.; Banks, J. L.; Murphy, R. B.; Halgren, T. A.; Klicic, J. J.; Mainz, D. T.; Repasky, M. P.; Knoll, E. H.; Shelley, M.; Perry, J. K.; Shaw, D. E.; Francis, P.; Shenkin, P. S. Glide: A new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy. *J. Med. Chem.* **2004**, *47* (7), 1739−1749.
(61) Predregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; Duchesnay, E. Scikit-learn: Machine Learning in Python. *JMLR* **2011**, *12*, 2825−2830.

https://doi.org/10.1021/acs.jpclett.1c00045
J. Phys. Chem. Lett. 2021, 12, 3724−3732