1, Identification of Allosteric Residues with Reversed Allosteric Effect (RAE).

Reversed allosteric effect (RAE) reflects response of a single residue against orthosteric perturbation. Those with strong RAE are identified as allosteric residues. RAE is mathematically defined as the change of residue-residue interactions inside a pocket between *apo* state and orthosteric ligand bound (*holo*) state, and was previously calculated using MM/GBSA decomposition along conformation ensemble generated by all-atom MD simulation (1,2). The main effort here, is to simplify the calculation.

Firstly, the input protein structure (binding orthosteric ligand) is reduced to coarse-grained (CG) models for the two states. In *apo* state, CG sites are placed at C$_\alpha$s of the protein. In *holo* state, atoms of orthosteric ligand are also turned into CG sites. For both systems, the potential energy function (3) is:

$$E = \sum_{i=0}^{n} \sum_{j=i+1}^{n} \frac{1}{2} \cdot k_{ij} \cdot \left(d_{ij} - d_{ij}^{0}\right)^{2} \cdot H\left(r_c - d_{ij}^{0}\right) \quad (1)$$

n is the number of CG sites. d_{ij} is distance between site *i* and *j*. d_{ij}^0 is the initial distance between site *i* and *j*. $H(x)$ is a Heaviside function. It equals 0 when $x \leq 0$ and 1 when $x > 0$. k_{ij} is elastic constant between site *i* and *j* and is set as 1 kcal/(mol·Å). The cutoff distance r_c is set as 12 Å.

Then we can calculate the interaction between residue *i* to all other residues in candidate pocket P (detected by FPocket (4)) according to eqn. (1):

$$I_i^s = \sum_{\substack{j \in P \\ j \neq i}} \frac{1}{4} \cdot k_{ij} \cdot \left(\langle d_{ij} \rangle^s - d_{ij}^{0}\right)^{2} \quad (2)$$

An additional one-second is multiplied here to avoid repeated calculation. Heaviside function is omitted here since every two residues in the pocket could have interactions. $<d_{ij}>^s$ is the average distance between site *i* and *j*. *s* defines the state and is either *apo* or *holo*. $<d_{ij}>^s$ is calculated by NMA. In brief, eigenvectors with corresponding eigenvalues are solved from the Hessian matrix of eqn. (1). These data describe the oscillations (also known as normal modes) from initial position with their frequencies for each CG site (or residue). Suppose for residue *i*, its displacement due to any of the normal modes obeys a normal distribution, we could describe the position ensemble of this residue at state *s* by:

$$R_i^s = R_i^0 + sf \sum_{p=1}^{q} X_p \cdot \lambda_p{}^{-0.5} \cdot u_p \quad (3)$$

, where R_i^0 describes the initial position of residue *i*, *q* is the number of modes used and is set to 100 here. λ_p and u_p are the eigenvalues and eigenvectors of mode *p*. *sf* is a scaling factor, ensuring the root mean square distance of the whole protein is 1 Å. X_p is a random variable obeying normal distribution. Therefore, $<d_{ij}>^s$ is written as:

$$\langle d_{ij} \rangle^s = \langle \left| R_i^s - R_j^s \right| \rangle \quad (4)$$

Unluckily, this expectation could not be analytically solved. Therefore, we solved this approach with a Monto-Carlo approach with 2000 samples. We found, though final results are not identical in 50 runs, the deviations are negligible, and it will not change the final prediction of allosteric sites.

Finally, RAE of residue i could be written as:

$$RAE_i = I_i^{holo} - I_i^{apo} \quad (5)$$

Normal mode analysis and pose sampling is done by Prody (5) package.

2, Recognition of Allosteric Sites.

2.1, Model Consrtruction.

AlloReverse identifies allosteric sites by discriminating whether a pocket-like region is allosteric or not using AdaBoost, a machine learning (ML) technique. AdaBoost is an ensemble learning framework using a serial of base models (6). The ith model is trained to patch the mistakes by the (i-1)th model. The model applies hydrophobicity, flexibility and RAE of pocket as input, and outputs judgement with prediction confidence. All three features are standardized according to protein. One-layer decision tree is applied as the base model. The model was trained on a dataset of 134 proteins and tested on a dataset of 58 proteins. In training set, 2431 pocket like regions are detected, and 208 of which with at least 10% overlap of recorded allosteric ligand are labeled as “allosteric site”, while the rest are labeled as “other site”. Due to great imbalance (about 1:10) between the two sites, oversampling method SMOTE was applied in training. Best super-parameters, including number of classifiers and learning rate, were grid-searched based on 5-fold cross validation on training set, where the average Matthews correlation coefficient (MCC) was used for index. The model is constructed by imblearn and scikit-learn toolkit (7).

2.2, Calculation of Input Features.

Hydrophobicity of pocket. This term is calculated with Fpocket. We would introduce how Fpocket detect pockets first. Fpocket first assigns alpha spheres in protein, which is a sphere decided by any 4 atoms in protein but with a restriction of radius range. Each alpha sphere decides a minor hole on protein. These holes are then clustered into several pockets on protein surface. In Fpocket, if more than 2 atoms, for decision of an alpha sphere, are polar atoms, then this alpha sphere is a polar alpha sphere. In Pocket P, we could calculate, for each polar alpha sphere, the number of other alpha spheres having overlap with current alpha sphere (n_{neigh}). Then hydrophobicity of pocket P, or “Mean local hydrophobic density” in Fpocket, is calculated as the average of n_{neigh} in pocket P.

Flexibility of pocket. This term is also calculated with Fpocket. Fpocket would define the atoms to form the pocket. The average B-factor among these atoms is then calculated. The value is finally normalized among all detected pockets on the protein surface.

RAE of pocket. RAE of the pocket is defined as the sum of RAE of residues in the pocket.

2.3, Empirical Adjustment After Prediction.

Prediction confidence is an output in AlloReverse by function “predict_proba” in scikit-learn toolkit. However, this value is not calibrated. We calibrate this value with the following relationship.

$$conf^{\Theta} = \min(2.5 * conf - 0.75, 1) \quad (6)$$

Due to oversampling applied, the model tends to give more positive predictions than reality. Therefore, we did the following adjustments. Pockets having more than 10% overlap with the orthosteric ligand is not a predicted allosteric site. In cases where above 70% pockets on protein are predicted to be allosteric sites, only sites in top 3 predicting confidence are selected as final prediction.

3, Prediction of Hierarchical Regulation Pathways.

Suppose the *holo* CG model is a graph and CG sites serve as nodes, regulation pathway toward pocket P is defined as the shortest route from the most central node of orthosteric ligand to the residue in pocket P with the highest RAE. The distance between every 2 nodes is defined as followed:

$$D_{ij} = \begin{cases} \dfrac{1}{corr_{ij}} & d_{ij}^{0} \leq 8\ \text{Å} \\ +\infty & d_{ij}^{0} > 8\ \text{Å} \end{cases} \quad (10)$$

Here, d_{ij}^{0} is the initial distance between site *i* and *j*; $corr_{ij}$ is the mean of Pearson correlation coefficient between node i and j at any direction or Top 100 modes. The shortest route is solved by Dijkstra algorithm (10).

4, Evaluation of Site-Site Couplings.

If pathway towards pocket P and Q each makes a set of residues named W_P and W_Q, then the site-site coupling score of pocket P by pocket Q is defined as:

$$c_{Q\to P} = \frac{\|W_P \cap W_Q\|}{\|W_P\|} \quad (11)$$

namely, the ratio of residues of pathway towards P shared by pathway towards Q. It could be seen that site-site coupling is asymmetry.

Reference

1. Zhang, Q., Chen, Y., Ni, D., Huang, Z., Wei, J., Feng, L., Su, J.-C., Wei, Y., Ning, S., Yang, X. *et al.* (2022) Targeting a cryptic allosteric site of SIRT6 with small-molecule inhibitors that inhibit the migration of pancreatic cancer cells. *Acta. Pharm. Sin. B*, **12**, 876-889.

2. Ni, D., Wei, J., He, X., Rehman, A.U., Li, X., Qiu, Y., Pu, J., Lu, S. and Zhang, J. (2020) Discovery of cryptic allosteric sites using reversed allosteric communication by a combined computational and experimental strategy. *Chem. Sci.*, **12**, 464-476.
3. Bahar, I. and Rader, A.J. (2005) Coarse-grained normal mode analysis in structural biology. *Curr. Opin. Struct. Biol.*, **15**, 586-592.
4. Le Guilloux, V., Schmidtke, P. and Tuffery, P. (2009) Fpocket: An open source platform for ligand pocket detection. *BMC Bioinformatics*, **10**, 168.
5. Zhang, S., Krieger, J.M., Zhang, Y., Kaya, C., Kaynak, B., Mikulska-Ruminska, K., Doruker, P., Li, H.C. and Bahar, I. (2021) ProDy 2.0: increased scale and scope after 10 years of protein dynamics modelling with Python. *Bioinformatics*, **37**, 3657-3659.
6. Freund, Y. and Schapire, R.E. (1997) A decision-theoretic generalization of on-line learning and an application to boosting. *J. Comput. Syst. Sci.*, **55**, 119-139.
7. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V. *et al.* (2011) Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.*, **12**, 2825-2830.
8. Huang, W., Wang, G., Shen, Q., Liu, X., Lu, S., Geng, L., Huang, Z. and Zhang, J. (2015) ASBench: benchmarking sets for allosteric discovery. *Bioinformatics*, **31**, 2598-2600.
9. Liu, X., Lu, S., Song, K., Shen, Q., Ni, D., Li, Q., He, X., Zhang, H., Wang, Q., Chen, Y. *et al.* (2020) Unraveling allosteric landscapes of allosterome with ASD. *Nucleic Acids Res.*, **48**, 394-401.
10. Abdelghany, H.M., Zaki, F.W. and Ashour, M.M. (2022) Modified Dijkstra Shortest Path Algorithm for SD Networks. *Int. J. Electr. Comput.*, **13**, 203-208.