HHS Public Access

Author manuscript

J Chem Inf Model. Author manuscript; available in PMC 2019 September 24.

Published in final edited form as:

J Chem Inf Model. 2018 September 24; 58(9): 1915–1925. doi:10.1021/acs.jcim.8b00314.

Efficiency of Stratification for Ensemble Docking using Reduced Ensembles

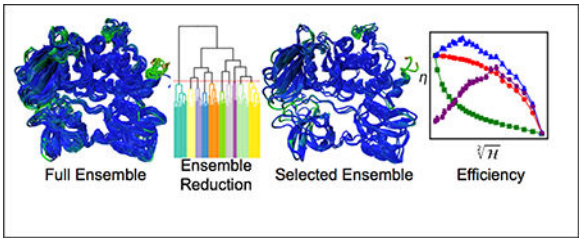
Bing Xie, **John D. Clark**, and **David D. L. Minh***

Department of Chemistry, Illinois Institute of Technology, Chicago, IL 60616, USA

Abstract

Molecular docking can account for receptor flexibility by combining the docking score over multiple rigid receptor conformations, such as snapshots from a molecular dynamics simulation. Here, we evaluate a number of common snapshot selection strategies using a quality metric from stratified sampling, the efficiency of stratification, which compares the variance of a selection strategy to simple random sampling. We also extend the metric to estimators of exponential averages (which involve an exponential transformation, averaging, and inverse transformation) and minima. For docking sets of over five hundred ligands to four different proteins of varying flexibility, we observe that for estimating ensemble averages and exponential averages, many clustering algorithms have similar performance trends: for few snapshots (less than 25), medoids are the most efficient while for a larger number, optimal (the allocation that minimizes the variance) and proportional (to the size of each cluster) allocation become more efficient. Proportional allocation appears to be the most consistently efficient for estimating minima.

Graphical Abstract



Introduction

Molecular docking is widely used to virtually screen large chemical libraries against biological targets to identify potential chemical probes and drug leads. Although it is clearly established that biological macromolecules are flexible and that their conformational

*To whom correspondence should be addressed dminh@iit.edu.

Supporting Information Available

Crystal structures and radii used in binding site definitions. Figures for: eigenvalues from principal components analysis for the four systems; stratification efficiency of QR factorization; comparison of linkage algorithms for archetypal senatorial and proportional allocation; and stratification efficiency trends for different types of distances. Tables for area under the $\eta(\sqrt{H})$ curves.

This material is available free of charge via the Internet at http://pubs.acs.org/.

equilibria may be influenced by ligand binding, docking programs, in order to be fast, often treat them as rigid.

One important strategy to ameliorate this approximation is ensemble docking, where docking is performed to multiple rigid receptor conformations[1–3]. These receptor conformations may be obtained in many ways, including from crystal structures[4,5], normal modes analysis[6], or molecular dynamics (MD). The latter can be performed with a receptor by itself or complexed to a ligand — this is known as the relaxed complex method[7–9] — or with enhanced sampling methods including replica exchange[10] and accelerated[11] MD, or even with a virtual ligand[12]. The overall score for the receptor-ligand pair is often based on the average docking score across the ensemble of receptor conformations[13–17]. Another common alternative is to use the minimum docking score of the ensemble[11,18].

In many cases, ensemble docking can be linked to standard binding free energies through implicit ligand theory (ILT)[19]. In ILT, coordinates of the entire complex, r_{RL}, are partitioned into internal coordinates of the receptor, r_R, and ligand, r_L, and external coordinates, ξ, that specify their position and orientation with respect to one another. The standard binding free energy between a receptor and ligand, ΔG°, has been shown to be an exponential average,

$$\Delta G^\circ = -\beta^{-1}\ln\left\langle e^{-\beta B(r_R)}\right\rangle_R^{r_R} + \Delta G_\xi, \quad (1)$$

of the binding potential of mean force (BPMF), $B(r_R)$, the binding free energy between a *flexible* ligand and *rigid* receptor configuration, r_R. $\beta = (k_B T)^{-1}$ is the inverse of Boltzmann's constant multiplied by the temperature. The angled brackets $\langle\ldots\rangle_R^{r_R}$ denote an average of receptor configurations over the ensemble of the receptor by itself without bound ligand (the apo ensemble). ΔG_ξ is the free energy of confining the external degrees of freedom into the binding site. A BPMF can also be expressed as an exponential average of the receptor-ligand interaction in implicit solvent, $\Psi(r_{RL})$, the difference between the energy of the complex and of the separated receptor and ligand. In physics-based scoring functions, the molecular docking score is usually based on $\Psi(r_{RL})$, so we will use this symbol to designate the docking score. Because exponential averages are typically dominated by low values of the exponent, the minimum $\Psi(r_{RL})$ — the typical goal of molecular docking — can be regarded as a *dominant state* approximation to the BPMF.

Based on this dominant state approximation, the most common combining rules in ensemble docking can be interpreted in terms of Equation 1. The ensemble average docking score is the first order in a cumulant expansion[20]. The minimum docking score can be treated as a dominant state approximation to ΔG°. In general, virtual screening based on the average and minimum docking score will identify different classes of ligands as strong binders. The average will identify ligands that bind tightly to all apo conformations and the minimum will identify those that bind tightly to a smaller subset. Because the former has a larger entropy, it can achieve the same or even greater binding affinity than the latter. The minimum may perform better in retrospective screens[11,18] because it is unnecessary that known ligands

(which are unlikely to have been discovered by the same virtual screening protocol) will adopt the strategy of binding to all apo configurations.

Of course, it is also feasible to use the exponential average in Equation 1 as a combining rule. To our knowledge, this has only been attempted by Nunes-Alves and Arantes[21] and by us[22]. For ligands of T4 lysozyme, neither study found that using exponential average was a better combining rule than the average or minimum. In our study, the different combining rules resulted in nearly equal performance in classifying molecules as active or inactive against the target (Figure 11 of Xie et al.[22]). In theirs, the dominant state approximation resulted in a smaller root mean square deviation (RMSD) and higher coefficient of determination, R^2 (Table 4 of Nunes-Alves and Arantes[21]). However, for human immunodeficiency virus reverse transcriptase and human FK506 binding protein 12, they found that the exponential average had a smaller RMSD, higher R^2, and smaller E_{max} (maximum difference between predicted and experimental values). These results suggest that the loss of entropy upon binding is relatively consistent among the known ligands to the artificial binding cavity in the T4 lysozyme L99A and L99A/M102Q mutants, which are largely small and rigid. However, differences in the entropy of binding are more important for the larger ligands to the pharmaceutically relevant drug targets in their study. Comparing the virtual screening performance of different combining rules for ensemble docking is an area worthy of further investigation, but outside the scope of the present paper.

One noteworthy aspect of ILT is that receptor conformations must be drawn from the apo ensemble or from a statistical distribution that can be appropriately reweighed. After all, Equation 1 is an average *over the apo ensemble*. Recently, we have also derived a version of ILT that enables *relative* binding free energy calculations based on an average over a holo ensemble.[23] The criterion for applying Equation 1 or the new expression from Nguyen and Minh[23] are satisfied for ensemble docking to MD trajectories, but using these expressions for a set of crystallographic structures is not statistically rigorous.

The most effective receptor sampling strategy is dependent on the nature of the binding process. The strategy of simulating the apo ensemble is simple and likely effective for ligands that bind by conformation selection, such that holo ensembles are similar to the apo ensemble. On the other hand, simulating a holo ensemble may be a useful strategy if binding occurs by induced fit, such that holo ensembles significantly differ the apo ensemble. In Xie et al.[22], we pursued yet another strategy, drawing and reweighing receptor configurations from a series of thermodynamic states in an alchemical pathway between apo and holo ensembles with six different reference ligands. Other strategies for generating holo-like receptor conformations, e.g. enhanced sampling algorithms, are an active area of research.

Once a suitable ensemble is generated, the computational cost of calculating ensemble averages can be decreased by using ensemble reduction techniques. Ensemble reduction techniques are useful because while simulation methods can generate a large number of configurations, it is computationally expensive to dock to all of them. Instead, docking can be performed to a representative subset[9]. The rationale behind selecting a representative subset is that when receptor configurations are similar, especially for atoms near the binding site, they are likely to have similar docking scores. Performing an independent docking

calculation for each configuration may be redundant, offering only a modest increase in accuracy for the increased computational expense. Thus, docking to a reduced ensemble can lead to significant computational savings.

Several methods for ensemble reduction have been developed. The most popular approach is to group the initial structures into clusters of similar conformations and to select a representative from each cluster. The most common way to measure the distance between conformations is the RMSD between α-carbons or between heavy atoms in the binding site (see the section on RMSD-based clustering in Amaro and Li[1]). Other distances based on structural properties of the binding cavity[24] and the occupancy of points on a three-dimensional grid[25,26] have also been devised. Besides clustering, ensemble reduction methods based on QR factorization[13] and maximizing differences between receptor-ligand interaction properties[27] have also been considered.

In addition to (or instead of) structure-based distances, ensemble reduction can also use experimental information about whether or not molecules bind to the receptor[6,28–33]. With these approaches, members of a training set are docked to all potential members of a reduced ensemble. The reduced ensemble is then selected based on which structure or combination of structures leads to the best discrimination between known active and inactive molecules. As with all knowledge-based strategies, these approaches to ensemble reduction are dependent on having a suitable training set, may not be applicable outside the training set (e.g. if a new ligand binds to a different pocket[34] or in a different way compared to training set ligands), and are difficult to interpret (e.g. it is unclear why particular structures are chosen.)

While ensemble reduction methods have been evaluated in a number of ways, none are directly related to the precision of ensemble averages. One class of quality metric is based on structural consistency[24,35,36], e.g. whether members of a cluster are closer to each other than to members of other clusters. Another class is based on virtual screening performance for a training set[6,28–33]. While improved virtual screening performance is the ultimate goal, it is unclear whether the right answers are obtained for the right reasons; ensemble reduction trained on discriminating known active and inactive molecules may not actually improve the precision of calculated of ensemble averages. Moreover, evaluating ensemble docking based on virtual screening performance does not point to a clear way to improve estimates. Lastly, Amaro et al.[9,13] compared the ensemble average from a large ensemble (400 structures) to the average from a representative ensemble (33 structures). The calculation provided anecdotal evidence that a reduced ensemble may be suitable for calculating the average docking score of a large ensemble, but did not provide information about statistical precision.

The key insight of this paper is that ensemble reduction methods are an application of a well-established statistical variance reduction technique, *stratified sampling*[37]. In stratified sampling, a population is broken up into strata in which members have similar characteristics. The average over the entire population is calculated as a weighted average of averages for each strata, with the weight being the size of the strata. Averages within each strata are usually but not necessarily obtained by simple random sampling *within the strata*.

There are a number of schemes to allocate samples to specific strata. If the number of samples is proportional to the size of the strata, this is referred to as *proportional allocation*. In *optimal allocation*, samples are allocated in a way that minimizes the variance of the final estimate. While optimal allocation yields more precise estimates, it is impractical because it requires prior knowledge of the variance of each stratum to calculate the number of samples to draw from each; obtaining this variance requires performing calculations on every data point. In ensemble docking, the most established procedure is to use the same number of samples within each strata. To our knowledge, this scheme has no established name in the context of stratified sampling. We will refer to it as *senatorial allocation*, after the Senate chamber in the United States legislature. Each state in the United States of America is allocated a fixed number of senators regardless of its population. In contrast, the number of congresspersons allocated to each state in the other legislative chamber, the House of Representatives, is an example of proportional allocation because it is (roughly) proportional to its population.

The recognition that ensemble reduction methods are stratified sampling is not merely a matter of taxonomy; it allows us to apply an established tool to assess the efficiency of stratified sampling to ensemble reduction methods: the *efficiency of stratification*, η. The efficiency of stratification compares the variance of averages obtained by stratified sampling to those based on random sampling. In this paper, we develop ways to analyze the efficiency of stratification for senatorial allocation. We also develop expressions to evaluate η for calculating exponential averages such as Equation 1. We then analyze molecular docking scores between libraries of known ligands and MD trajectories of four proteins — Abl kinase, cruzipain, dihydrofolate reductase (DHFR), and estrogen receptor α (ERα) — comparing the efficiency of stratification for senatorial, proportional, and optimal allocation with a number of different ensemble reduction techniques.

Theory

Efficiency of Stratification for Ensemble Averages

In stratified sampling, the statistical estimator for the ensemble average of x is,

$$\bar{x} = \sum_{h}^{H} \left(\frac{N_h}{N} \right) \bar{x}_h, \quad (2)$$

where N is the total population size, where $h \in \{1, \ldots, H\}$ is an index for the strata, N_h is the number of samples in strata h, and $\bar{x}_h$ is the estimate for the expectation of x within strata h.

The variance of Equation 2 is based on the variance of a linear combination of independent random variables $\bar{x}_h$,

$$\sigma^2_{strat}[\bar{x}] = \sum_{h}^{H} \left(\frac{N_h}{N} \right)^2 \sigma^2[\bar{x}_h]. \quad (3)$$

On the other hand, for simple random sampling from the entire population, if the finite population correction is neglected, then the variance of the sample mean based on n samples is,

$$\sigma^2_{simple}[\bar{x}] = \frac{1}{n}\sigma^2[x]\,. \quad (4)$$

The efficiency of stratification is based on the ratio of $\sigma_{strat}[\bar{x}]$ from Equation 3 to $\sigma_{simple}[\bar{x}]$ from Equation 4,

$$\eta_{strat} = \frac{n}{\sigma^2[x]}\sum_h^H \left(\frac{N_h}{N}\right)^2 \sigma^2\left[\bar{x}_h\right]. \quad (5)$$

While Equation 5 includes the total number of samples, n, for simple random sampling within each strata, it usually cancels out with another n from $\sigma^2\left[\bar{x}_h\right]$ such that the efficiency of stratification is independent of the sample size.

If the finite population correction – which accounts for the fact that samples are drawn without replacement from a finite population instead of with replacement from an infinite population – is included, then the efficiency of stratification is dependent on sample size. The finite population correction to Equation 4 is $\left(\frac{N-n}{N-1}\right)$. For simple random sampling within each strata, the variance $\sigma^2\left[\bar{x}_h\right]$ also includes $\left(\frac{N_h - n_h}{N_h - 1}\right)$, which does not cancel with the finite population correction for simple random sampling from the whole population. For simplicity, we will neglect the finite population correction; we observed that our qualitative efficiency trends are very similar with and without the correction.

If η_{strat} is less than one, stratified sampling is superior to simple random sampling from the entire population. The efficiency of stratification can also be thought of as a ratio of the effective sample size. For example, $\eta_{strat} = 0.5$ implies that when using stratified sampling instead of simple random sampling, the same precision can be achieved with half the number of samples. Alternately, stratified sampling has twice the effective sample size of simple random sampling. On the other hand, if η_{strat} is equal to one then stratified sampling and simple random sampling are equally efficient. Lastly, if η_{strat} is greater than one then stratified sampling is less efficient.

If every data point is its own strata, then every $\sigma^2\left[\bar{x}_h\right]$ is zero and the stratification efficiency is zero. If there is a single stratum and samples are drawn by simple random sampling, then η_{strat} is always equal to one. In other situations the efficiency of stratification will depend on the allocation algorithm.

Senatorial Allocation

In senatorial allocation, there are an equal number of samples from each strata. For n total samples, the number of samples drawn from strata h is n/H. If samples within each strata are drawn by simple random sampling, the variance of the mean within each strata is, $\sigma^2[\bar{x}_h] = \frac{H}{n}\sigma^2[x_h]$, and the efficiency of stratification is,

$$\eta_{sen} = \frac{H}{\sigma^2[x]} \sum_h^H \left(\frac{N_h}{N}\right)^2 \sigma^2[x_h]. \quad (6)$$

However, the standard practice in ensemble docking is not simple random sampling within each strata. The calculated strata average is based on a non-random representative, such as the medoid of the strata. We will refer to this approach as *archetypal senatorial allocation*. In general, the variance of an estimator is defined as the expectation of the squared deviation of an estimate from its true value, $\sigma^2[\bar{x}_h] = \left\langle (\bar{x}_h - \langle x_h \rangle)^2 \right\rangle$. Thus if the estimate is constant, then the variance is simply $\sigma^2[\bar{x}_h] = (\bar{x}_h - \langle x_h \rangle)^2$. By substitution into Equation 5,

$$\eta^*_{sen} = \frac{H}{\sigma^2[x]} \sum_h^H \left(\frac{N_h}{N}\right)^2 (\bar{x}_h - \langle x_h \rangle)^2. \quad (7)$$

We have assumed that there is one sample per strata such that the total number of samples n is equal to the number of strata H.

Proportional Allocation

In proportional allocation, the number of samples drawn from strata h is proportional to N_h. For n total samples, the number of samples drawn from strata h is nN_h/N. If samples within each strata are drawn by simple random sampling, the variance of the mean within each strata is $\sigma^2[\bar{x}_h] = \frac{N}{nN_h}\sigma^2[x_h]$. By substitution of this variance expression into Equation 5, the efficiency of stratification is,

$$\eta_{\propto} = \frac{1}{\sigma^2[x]} \sum_h^H \left(\frac{N_h}{N}\right) \sigma^2[x_h]. \quad (8)$$

Optimal Allocation

Optimal allocation is derived by minimizing the variance of the estimator with respect to the number of samples allocated to strata h. For simple random sampling of n_h samples from strata h, then the variance of the estimate is, $\sigma^2[\bar{x}] = \sum_h^H \left(\frac{N_h}{N}\right)^2 \frac{1}{n_h}\sigma^2[x_h]$. Using Lagrange

multipliers, we would like to minimize this variance subject to the constraint that $g(n_h) = \sum_h^H n_h - n = 0$. The total differential of the constrained variance is,

$$d\left(\sigma^2[\bar{x}] - \alpha g\right) = \sum_h^H \left[\frac{\partial \sigma^2[\bar{x}]}{\partial n_h} - \alpha \frac{\partial g}{\partial n_h}\right] dn_h \quad (9)$$
$$= \sum_h^H \left[-\left(\frac{N_h}{N}\right)^2 \frac{1}{n_h^2}\sigma^2[x_h] - \alpha\right] dn_h .$$

Setting each term in the sum to zero, yields $n_h \propto N_h \sigma[x_h]$. Hence, for n total samples, $n_h = n\left(\frac{N_h \sigma[x_h]}{\sum_k N_k \sigma[x_k]}\right)$. With this allocation, the variance is,

$$\sigma^2[\bar{x}] = \frac{1}{n}\sum_h^H \left(\frac{N_h}{N}\right)^2 \frac{\sum_k N_k \sigma[x_k]}{N_h \sigma[x_h]} \sigma^2[x_h] \quad (10)$$
$$= \frac{1}{n}\left(\sum_h^H \frac{N_h}{N}\sigma[x_h]\right)^2 ,$$

and the efficiency of stratification is,

$$\eta_{opt} = \frac{1}{\sigma^2[x]}\left(\sum_h^H \frac{N_h}{N}\sigma[x_h]\right)^2 . \quad (11)$$

The efficiency of stratification for proportional and optimal allocation are well-established[37] and are included to introduce important notation and so that the paper is better self-contained.

Efficiency of Stratification for Exponential Averages

In certain situations, it is of greater interest to evaluate an exponential average of the form,

$$F = -\beta^{-1} \ln\left\langle e^{-\beta x}\right\rangle, \quad (12)$$

than the expectation value of x. β is a constant. For convenience, let us define $y = e^{-\beta x}$ and y_h to be the expectation value of y within strata h.

The statistical estimator for Equation 12 based on stratified sampling is,

$$\bar{F} = -\beta^{-1} \ln \sum_h^H \left(\frac{N_h}{N}\right) \bar{y}_h . \quad (13)$$

Using error propagation based on a first-order Taylor series expansion, its variance is,

$$\sigma^2_{strat}[\bar{F}] = \frac{\beta^{-2}\sigma^2\left[\sum_h^H \left(\frac{N_h}{N}\right)\bar{y}_h\right]}{\left[\sum_h^H \left(\frac{N_h}{N}\right)y_h\right]^2} = \frac{\beta^{-2}\sum_h^H \left(\frac{N_h}{N}\right)^2 \sigma^2[\bar{y}_h]}{e^{-2\beta F}}. \quad (14)$$

The term in the denominator is based on evaluating the partial derivative that appears in the expansion at the mean, or the true value of y_h. With simple random sampling, in contrast, the statistical estimator for Equation 12 is,

$$\bar{F} = -\beta^{-1}\ln\left(\frac{1}{N}\sum_n y_n\right). \quad (15)$$

Using the same error propagation procedure, its statistical variance is,

$$\sigma^2_{simple}[\bar{F}] = \frac{\beta^{-2}\sigma^2[\bar{y}]}{y^2} = \frac{\beta^{-2}\sigma^2[y]}{ne^{-2\beta F}}. \quad (16)$$

The efficiency of stratification is based on the ratio of $\sigma_{strat}[\bar{F}]$ from Equation 14 to $\sigma_{simple}[\bar{F}]$ from Equation 16,

$$\eta_{strat} = \frac{n}{\sigma^2[y]}\sum_h^H \left(\frac{N_h}{N}\right)^2 \sigma^2[\bar{y}_h]. \quad (17)$$

This expression is exactly analogous to Equation 5 except that x has been replaced by y. Hence, expressions for efficiency of stratification for senatorial, proportional, and optimal allocation are also analogous to Equations 6, 8, and 11 above and will not be reproduced here.

Efficiency of Stratification for Minima

Now consider the efficiency of stratification for estimating the minimum. For convenience, let us define $z = \min(x)$. The variance of an estimator for z is $\sigma^2[\bar{z}] = \left\langle(\bar{z} - z)^2\right\rangle$. There is no central limit theorem that describes how the variance of the estimated minimum changes

with sample size. However, this variance can be estimated by bootstrapping. To obtain the variance for n snapshots, we repeatedly resampled n snapshots from the population with replacement, used the minimum from each resampling as the point estimate, and then took the sample variance of the point estimates. For archetypal senatorial allocation, the variance is the squared difference between the estimated and true minima.

Computational Methods

System Preparation and Molecular Docking

Structures of abl kinase, cruzipain, DHFR, and ERα were downloaded from the Protein Data Bank (PDB IDs (chain) were 1opj(A), 1me4(A), 1j3j(A), and 1x7e(A), respectively). Protein protonation states were predicted with pdb2pqr 1.9.0[38] at a pH of 7.0. Using AmberTools 14[39], parameter/topology files were prepared based on the AMBER ff14SB force field[40]. Solvent was treated using the generalized Born model 2 from Onufriev, Bashford, and Case (OBC2)[41].

Using OpenMM 6.2[42], the structures were minimized and then simulated with Langevin dynamics at 300 K using a time step of 2 fs. Snapshots were stored every 1 ps. The first 20 ns were discarded as equilibration. 800 snapshots from the last 80 ns were used for production. The snapshots were split into two sets of 400 based on every other saved simulation snapshot.

SMILES strings[43] for sets of up to 520 (520, 518, 518, 520, respectively) known ligands for each receptor were obtained from the BindingDB[44] (http://www.bindingdb.org/). Balloon[45] was used to generate 3D models from the SMILES strings and UCSF Chimera[46] was used to protonate ligands and to calculate AM1-BCC partial charges[47,48]. AmberTools 14[39] was used to prepare parameter/topology files based on the Generalized Amber Force Field.[49]

The binding site of each receptor was defined based on sets of homologous crystal structures. First, a UniProt[50] template sequence for each protein was selected (Table 1). Next, MODELLER 9.18[51] was used to identify crystal structures with sequences with at least a minimum sequence identity (specified in Table 1) compared with the template. A reference chain was selected arbitrarily. ProDy 1.8.2[52] was then used to align each structure by minimizing the α carbon RMSD relative to a reference chain. The center-of-mass (COM) was then calculated for every ligand. The binding site center was defined as the mid-point of the maximum COM and minimum COM. The binding site radius was defined by rounding the maximum distance from the site center to a ligand COM up to the nearest Ångstrom. Because the ligands that have been co-crystallized with some receptors are often very similar, some of the binding sites were manually expanded. PDB IDs of homologous crystal structures and binding site radii are described in Section S1 of the Supporting Information.

Each of the ligands was then docked to the 800 snapshots using UCSF DOCK 6[53].

Sphgen was run with default parameters: dotlim = 0.0, radmax = 4.0 Å, and radmin = 1.4 Å. Spheres for docking were selected from clusters with spheres within the binding site, as

defined in the previous paragraph. The docking grid was defined with the binding site center as its center and the edge length,

$$l = 2 \times \text{ceil}\left(r + d_{max,\,atom}\right), \quad (18)$$

where the function ceil(·) denotes rounding up to the nearest integer, r is the binding site radius, and $d_{max,atom}$ is the maximum distance between crystallographic ligand atoms and the center. The grid spacing was 0.25. DOCK was run with 5000 maximum orientations, using internal energy with an exponent of 12, a flexible ligand, an a minimum_anchor_size of 40. Pruning was performed with clustering, 1000 maximum orientations, a clustering cutoff of 1000, and conformer score cutoff of 25.0. A bump filter was used with max_bumps_anchor = 12 and max_bumps_growth = 12. Final conformations were clustered with a root mean square deviation (RMSD) threshold of 2.0 Å. If no binding pose was obtained, the docking score was assumed to be zero.

Ensemble Reduction

Ensemble reduction was performed based on three types of distances:

1. The **RMSD** between: (a) α-carbons in the whole protein, (b) α-carbons in the binding site, and (c) heavy atoms in the binding site.

2. **Principal components analysis (PCA)**, a dimensionality reduction technique that describes motion in a different way from the RMSD. PCA was performed on the aligned coordinates of α-carbons in the whole protein. Because the eigenvalues significantly dropped off within the first six eigenvalues (Figure S1 in the Supporting Information), projections were made onto the first six eigenvectors of each snapshot. The distance matrix was based on the Euclidean distance,

$$d_{ij} = \sqrt{\sum_{k=1}^{6} \left(c_{ik} - c_{jk}\right)^2}, \quad (19)$$

where d_{ij} is the distance between snapshots i and j and c_{ik} is the projection of snapshot i onto eigenvector k.

3. **Occupancy fingerprints**, based on the occupancy of points spaced 0.25 Å apart on a three-dimensional grid in the binding site. They were marked as occupied if they were within the van der Waals radius of a receptor atom. These distances were calculated by minor modification to POVME 2.0[54]. For a pair of grids, M_{nm} represents the total number of points where the first grid has the value n and the second the value m. Distance matrices were based on the occupancy fingerprint overlap,

$$d_{overlap}ij = M_{10} + M_{01}, \quad (20)$$

Tanimoto-inspired similarity,

$$d_{tanimoto}ij = -log_2\left(\frac{M_{11}}{M_{11} + M_{10} + M_{01}}\right), \quad (21)$$

and Jaccard distance,

$$d_{jaccard}ij = \left(\frac{M_{10} + M_{01}}{M_{11} + M_{10} + M_{01}}\right). \quad (22)$$

Offutt et al.[25] also used the Tanimoto similarity and and Motta and Bonati[26] used Equation 22.

4. **The Q_H structural similarity measure**[55], which compares interatomic distances between structures. For a pair of protein structures with the exact same amino acid sequence, $Q_H \propto \sum_j e^{-\left(\Delta d_j/\sigma_j\right)^2}$, where j is the index for d_j, an interatomic distance between α carbons. Δd_j is the difference in this distance between two structures. The standard deviation σ_j that weights each difference is dependent on the index. Q_H is normalized as described in O'Donoghue and Luthey-Schulten[55]. It was calculated using the QR factorization module in VMD 1.9.7[56].

Before clustering, snapshots were aligned to the reference crystal structure by VMD 1.9.7[56]. For the RMSD between all α carbons in the whole protein, PCA, and the Q_H-QR calculations, structures were aligned according to all α carbons in the whole protein. In the other calculations, structures were aligned according to atoms in the binding site.

For most calculations, ensemble reduction was performed based on hierarchical clustering. Hierarchical clustering was performed using the scipy.cluster (http://www.scipy.org/) package using the different types of distances and with four different linkage criteria: single, complete, average, and weighted. In our analysis, each cluster was considered a strata.

We also considered ensemble reduction based on QR factorization[55], which was previously used to improve the relaxed complex method[9] and is implemented in VMD[56]. QR factorization is a well-established linear algebra in which a matrix A is decomposed into an orthogonal matrix Q and upper triangular matrix R such that $A = QR$. O'Donoghue and Luthey-Schulten[55] applied a multidimensional variant of QR factorization to order protein structures by increasing linear dependence. We used a Q_H threshold of 0.9, similar to the threshold of 0.86 previously used in the relaxed complex method[9]. After representative structures were determined, we calculated distances from each snapshot to each

representative snapshot. The cluster of each snapshot was assigned based on the nearest representative.

After clustering, the stratification efficiency was plotted as a function of $\sqrt{H}$. To summarize this plot, the area under the curve was computed using the trapezoidal rule, as implemented in the numpy.trapz function in numerical python[57] 1.13.3 (http://www.scipy.org/).

Our data and analysis scripts are available at https://github.com/bxie4/stratification_efficiency_data.git.

Results and Discussion

Distance matrices provide unique angles on trajectories of varying flexibility

Even when considering the same MD trajectories, different distance matrices are qualitatively unique from one another (Figure 1). Because elements of these matrices are ordered by time, they consist of blocks with correlated configurations along the diagonal. The precise boundaries of these blocks and the magnitude of off-diagonal blocks are key properties that distinguish one type of distance from another.

The most significant factor distinguishing our matrices is choice of atoms for alignment and distance calculations. Distances based on atoms from the entire protein, which appear on the left column of Figure 1, are more similar to each other than those based on atoms in the binding site, which appear on the right column of the corresponding figures. Considering Abl kinase, for example, matrices on the left side consist of three major diagonal blocks, but those on the right side appear to have four. Furthermore, according to the RMSD between all α carbons, the first fourth of the Abl kinase trajectory has little structural resemblance to latter portions of the trajectory. On the other hand, according to the RMSD between α carbons in the binding site, the structures that emerge after about half of the trajectory are not that different.

The Jaccard distance is unique because it is bounded. Whereas other types of distances have no upper bound, the Jaccard distance is bounded between 0 and 1. In the other panels, because we used the same color scale for all four proteins, the reddest colors are established by the largest observed distance in any of the systems. This makes on-diagonal blocks look particularly blue. In Figure 1d, the reddest color is 1 and blue only appears with *very* similar binding sites. For this reason, Figure 1d appears to (but actually does not) have less time correlation than the other panels.

The distance matrices reveal different levels of receptor flexibility in each of the four MD simulations. Cruzipain was comparatively rigid. The RMSD between binding site α carbons remained within 1.5 Å for all snapshots and within 0.6 Å for most snapshots. It also had the smallest range of Jaccard distances between occupancy fingerprints. According to the RMSD between all α carbons, Abl kinase had the largest range of overall flexibility. However, the Jaccard distance between occupancy fingerprints shows that the shape of the binding site remained fairly constant. Compared to cruzipain and Abl kinase, DHFR and ER

α had an intermediate range of RMSDs between all α carbons and a larger range of Jaccard distances between occupancy fingerprints.

The distance matrices could be used to argue that most of the MD simulations have not fully converged. Fully converged simulations would return to the same conformations multiple times. For our present purposes, however, it is not relevant whether the snapshots correctly represent the Boltzmann distribution and adequately represent the configuration space available to the protein. Instead, we are treating the sampled snapshots as our populations and assessing methods to estimate summary statistics for these populations based on a subset of snapshots.

Our summary statistics of interest are based on docking scores, which have different levels of variability (Figure 2). For most ligands, the docking scores of cruzipain and Abl kinase have a small standard deviation. In contrast, most ligands for DHFR have large range of docking scores. For ER α, the distribution of docking score standard deviations is bimodal. The small standard deviations of docking scores for cruzipain and Abl kinase ligands are consistent with the small maximal Jaccard distance between occupancy fingerprints. Likewise, broader range of standard deviations for DHFR and ER α reflect the larger maximal Jaccard distance.

Stratification efficiencies are sensitive to multiple factors

Ensemble reduction involves a number of choices including the type of distance, the clustering algorithm, and the allocation scheme. Each of these choices influences the efficiency of stratification in different ways. Tables S1, S2, and S3 in the Supporting Information show the area under the $\eta(\sqrt{H})$ curve for ensemble averages, exponential averages, and minima, respectively. In the following sections, we will first describe the influence of these choices for ensemble averages.

Allocation affects the shape of $\eta(\sqrt{H})$ curves

Each allocation scheme has a characteristic behavior of η as a function of the number of strata (Figure 3). Indeed, the choice of allocation scheme is the factor with the largest influence on the area under this curve (Table S1 in the Supporting Information). With archetypal senatorial allocation, the stratification efficiency exhibits a hump-shaped curve. This curve starts with a low value, approximately 0.1, and gradually increases until $H \approx 100$ before gradually decreasing to zero. Although the curves are noisier, the same qualitative trend is seen for archetypal senatorial allocation based on QR factorization (Figure S2 in the Supporting Information). With the other allocation schemes, η starts near 1. For optimal allocation, η quickly decreases for $H < 25$ and decreases more gradually for larger H. For proportional allocation, η gradually decreases as H increases. For senatorial allocation, η actually increases for low H and then gradually decreases. These qualitative trends held across all four systems studied.

Observed trends in stratification efficiency validate the most common practice in the field of ensemble docking. For small H ($H < 25$), the best algorithms for archetypal senatorial allocation have lower η and thereby outperform the other allocation schemes. In contrast, for

$H > 25$, η is lower for optimal allocation and increases between optimal, proportional, and senatorial allocation. The contrast between archetypal senatorial allocation and senatorial allocation based on random samples is especially stark. This contrast shows that, for calculating the average, the medoid is more suitable than an arbitrary representative. The low η of archetypal senatorial allocation shows that selecting a single representative from each cluster is a reasonable strategy for estimating ensemble averages with a small number of snapshots.

The results also suggest that for larger H, variance-based allocation may be beneficial. The difference between the efficiency of optimal and proportional allocation is substantial. Unfortunately, because it is impossible to determine the precise variance of docking scores without performing docking to all structures, optimal allocation is an idealization that is not practical to fully implement. However, a subset of allocated computational effort could be used to estimate the variance prior to allocating remaining computational effort. With such a procedure, it should be possible to achieve an efficiency between that of optimal and proportional allocation.

Linkage criteria also affect stratification efficiency

With hierarchical clustering methods, the efficiency of stratification is sensitive to the linkage criterion and specific to the allocation method (Table S1 in the Supporting Information). For optimal allocation, complete linkage has the lowest η, followed by weighted, average, and single linkage (Figure 4). As a number of clusters increases, the efficiency of weighted and average linkages become comparable to complete linkage. For archetypal allocation, most linkage criteria have similar efficiency (Figure 4). The exception is single linkage, which has the worst efficiency. In all systems except for DHFR, the area under the curve for $H < 25$ is the smallest for complete linkage. For proportional allocation, most linkage criteria have similar efficiency, but single linkage performs well for a large number of clusters (Figure S3 in the Supporting Information).

The type of distance is relevant for senatorial allocation

The stratification efficiency is sometimes but not always sensitive to the type of distance. For optimal and proportional allocation, the area under the η vs. $\sqrt{H}$ curve is fairly consistent across different types of distances (Table S1 in the Supporting Information). On the other hand, using different distances can lead to significant efficiency differences in senatorial allocation, both when the sample is the medoid or it is random. Because of its lower η, we will focus on archetypal senatorial allocation.

In archetypal senatorial allocation, no type of distance has completely consistent performance across all systems. However, distances based on occupancy fingerprints have similar performance and are among the best in all systems, especially for low H (Figure 5). The superior performance of these distances are most evident for cruzipain. It is least evident for DHFR, especially with larger H.

For DHFR, the relatively poor performance of the Jaccard distance between occupancy fingerprints for larger H may be due to the effects of electrostatics on docking scores.

Electrostatic effects have a large influence on ligand binding to DHFR; the enzyme possesses pockets of positive potential in its active site that enable it to bind negatively charged substrates.[58] Because occupancy fingerprints only consider short-range van der Waals interactions and neglect electrostatics, they may not work as well for clustering conformations with similar docking scores. In contrast, although ER α has similar flexibility and a larger maximum Jaccard distance (Figure 1), the Jaccard distance is effective because electrostatics have a smaller influence on ligand binding; the native ligand estrogen is a mostly hydrophobic hormone with no net charge.

Qualitative trends in $\eta(\sqrt{H})$ are similar for ensemble averages and exponential averages, but distinct for minima

Regardless of allocation scheme, the stratification efficiency of exponential averages is very similar to that of ensemble averages (see Figures 6 and S4 in the Supporting Information for archetypal senatorial allocation, and Figures 7 and S5 in the Supporting Information for proportional allocation, and Figure S6 in the Supporting Information for optimal allocation). The exponential average of the docking score appears to cluster in a similar way as its non-transformed counterpart. Thus, the same conclusions about the most efficient allocation scheme and other choices hold for both ensemble averages and exponential averages.

In contrast, stratification efficiency for minima follow different trends. With archetypal senatorial allocation, η starts between 0.5 and 1, depending on the system, and gradually decreases with increasing H. For some systems, the decrease is not monotonic and smooth. On the other hand, for proportional allocation, η starts at 1.0 and decreases very quickly with H. In some cases, it increases at high H. The increase at high H is likely due to the fact that simple random sampling becomes increasingly likely to draw from all strata, diminishing the benefits of stratified sampling.

Notably, efficient strategies for minima also contrast with the averages because of the best linkage criterion. For calculating ensemble averages and exponential averages, the single linkage criterion usually is the worst, with the largest area under the curve (Tables S1 and S2 in the Supporting Information). In contrast, for obtaining the minimum docking score, the single linkage criterion is usually the best. The most consistently efficient methods for obtaining a minimum docking score are based on hierarchical clustering with single linkage and either the RMSD between binding site α carbons or the Jaccard distance between occupancy fingerprints (Table S3 in the Supporting Information).

Conclusions and Future Directions

A tool from stratified sampling, the efficiency of stratification, was imported and used in the context of ensemble docking with reduced ensembles. We applied the efficiency of stratification for ensemble averages, developed theory pertaining to exponential averages, and implemented methods to estimate η for minima. A variety of ensemble reduction methods incorporating different types of distances and clustering algorithms were compared. We applied our analysis to four systems with varying flexibility.

For estimating ensemble averages, we found that the prevailing approach, archetypal senatorial allocation, was most efficient for small numbers of clusters. For a larger number of clusters, optimal and proportional allocation are more efficient. With optimal allocation, hierarchical clustering based on the complete linkage criterion was found to perform similarly or better than other clustering algorithms. For most systems, the Jaccard distance between occupancy fingerprints was one of the best performing types of distances with archetypal senatorial allocation.

While similar performance trends were observed for exponential averages, minima behave differently. For estimating minima, hierarchical clustering based on the single linkage criterion with a distance based on the Jaccard distance between occupancy fingerprints was found to be the most efficient for nearly any cluster size.

Based on their performance for these four systems, we cautiously recommend these clustering algorithms and allocation schemes in future ensemble docking calculations. Our recommendation is cautious because we have analyzed a limited number of systems and types of ensembles. Regarding the latter point, we have focused our analysis on apo ensembles and thereby only considered ligands that bind by conformational selection. Although we have no particular reason to think so, it could be possible that other ensemble reduction schemes are better suited for docking to other, e.g. ligand-bound, ensembles. Because docking to other types of ensembles may be useful strategy for ligands that bind by induced fit, analysis of the efficiency of stratification for these ensembles could be a fruitful future research direction. Beyond analyzing more systems and different types of ensembles, we also suggest that the efficiency of stratification should be used to assess other algorithms, existing and new, for ensemble reduction.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgement

Financial support for this research was provided by the National Institute of General Medical Sciences in the National Institutes of Health, under grant R15GM114781.

References

(1). Amaro RE; Li WW Emerging Methods for Ensemble-Based Virtual Screening. Curr. Top. Med. Chem 2010, 10, 3–13. [PubMed: 19929833]

(2). Feixas F; Lindert S; Sinko W; McCammon JA Exploring the Role of Receptor Flexibility in Structure-Based Drug Discovery. Biophys. Chem 2014, 186, 31–45. [PubMed: 24332165]

(3). Amaro RE; Baudry J; Chodera J; Demir Ö; McCammon JA; Miao Y; Smith JC Ensemble Docking in Drug Discovery. Biophys. J 2018, 114, 2271–2278. [PubMed: 29606412]

(4). Bottegoni G; Rocchia W; Rueda M; Abagyan R; Cavalli A Systematic Exploitation of Multiple Receptor Conformations for Virtual Ligand Screening. PLoS One 2011, 6, e18845. [PubMed: 21625529]

(5). Hou X; Li K; Yu X; Sun J.p.-; Fang H Protein Flexibility in Docking-Based Virtual Screening: Discovery of Novel Lymphoid-Specific Tyrosine Phosphatase Inhibitors Using Multiple Crystal Structures. J. Chem. Inf. Model 2015, 55, 1973–1983. [PubMed: 26360643]

(6). Moroy G; Sperandio O; Rielland S; Khemka S; Druart K; Goyal D; Perahia D; Miteva MA Sampling of Conformational Ensemble for Virtual Screening Using Molecular Dynamics Simulations and Normal Mode Analysis. Future Med. Chem 2015, 7, 2317–2331. [PubMed: 26599419]

(7). Lin JH; Perryman AL; Schames JR; McCammon JA The Relaxed Complex Method: Accommodating Receptor Flexibility for Drug Design with an Improved Scoring Scheme. Biopolymers 2003, 68, 47–62. [PubMed: 12579579]

(8). Lin JH; Perryman AL; Schames JR; McCammon JA Computational Drug Design Accommodating Receptor Flexibility: The Relaxed Complex Scheme. J. Am. Chem. Soc 2002, 124, 5632–5633. [PubMed: 12010024]

(9). Amaro RE; Baron R; McCammon JA An Improved Relaxed Complex Scheme for Receptor Flexibility in Computer-Aided Drug Design. J. Comput.-Aided Mol. Des 2008, 22, 693–705. [PubMed: 18196463]

(10). Buonfiglio R; Ferraro M; Falchi F; Cavalli A; Masetti M; Recanatini M Collecting and Assessing Human Lactate Dehydrogenase-A Conformations for Structure-Based Virtual Screening. J. Chem. Inf. Model 2013, 53, 2792–2797. [PubMed: 24138094]

(11). Miao Y; Goldfeld DA; Moo EV; Sexton PM; Christopoulos A; McCammon JA; Valant C Accelerated Structure-Based Design of Chemically Diverse Allosteric Modulators of a Muscarinic G Protein-Coupled Receptor. Proc. Natl. Acad. Sci. USA 2016, 113, E5675–E5684. [PubMed: 27601651]

(12). Xu M; Lill MA Significant Enhancement of Docking Sensitivity Using Implicit Ligand Sampling. J. Chem. Inf. Model 2011, 51, 693–706. [PubMed: 21375306]

(13). Amaro RE; Schnaufer A; Interthal H; Hol W; Stuart KD; McCammon JA Discovery of Drug-Like Inhibitors of an Essential RNA-Editing Ligase in Trypanosoma Brucei. Proc. Natl. Acad. Sci. USA 2008, 105, 17278–17283. [PubMed: 18981420]

(14). Durrant JD; Keränen H; Wilson BA; McCammon JA Computational Identification of Uncharacterized Cruzain Binding Sites. PLoS Negl. Trop. Dis 2010, 4, e676. [PubMed: 20485483]

(15). Chan AH; Wereszczynski J; Amer BR; Yi SW; Jung ME; McCammon JA; Clubb RT Discovery of Staphylococcus Aureus Sortase A Inhibitors Using Virtual Screening and the Relaxed Complex Scheme. Chem. Biol. Drug Des 2013, 82, 418–428. [PubMed: 23701677]

(16). Ochoa R; Watowich SJ; Flórez A; Mesa CV; Robledo SM; Muskus C Drug Search for Leishmaniasis: a Virtual Screening Approach by Grid Computing. J. Comput.-Aided Mol. Des 2016, 30, 541–552. [PubMed: 27438595]

(17). Hart KM; Moeder KE; Ho CM; Zimmerman MI; Frederick TE; Bowman GR Designing Small Molecules to Target Cryptic Pockets Yields Both Positive and Negative Allosteric Modulators. PLoS ONE 2017, 12, 1–13.

(18). Swift RV; Jusoh SA; Offutt TL; Li ES; Amaro RE Knowledge-Based Methods to Train and Optimize Virtual Screening Ensembles. J. Chem. Inf. Model 2016, 56, 830–842. [PubMed: 27097522]

(19). Minh DDL Implicit Ligand Theory: Rigorous Binding Free Energies and Thermodynamic Expectations From Molecular Docking. J. Chem. Phys 2012, 137, 104106. [PubMed: 22979849]

(20). Zwanzig R High-Temperature Equation of State by a Perturbation Method. I. Nonpolar Gases. J. Chem. Phys 1954, 22, 1420.

(21). Nunes-Alves A; Arantes GM Ligand-Receptor Affinities Computed by an Adapted Linear Interaction Model for Continuum Electrostatics and by Protein Conformational Averaging. J. Chem. Inf. Model 2014, 54, 2309–2319. [PubMed: 25076043]

(22). Xie B; Nguyen TH; Minh DDL Absolute Binding Free Energies Between T4 Lysozyme and 141 Small Molecules: Calculations Based on Multiple Rigid Receptor Configurations. J. Chem. Theory Comput 2017, 13, 2930–2944. [PubMed: 28430432]

(23). Nguyen TH; Minh DDL Implicit Ligand Theory for Relative Binding Free Energies. J. Chem. Phys 2018, 148, 104114. [PubMed: 29544299]

(24). De Paris R; Quevedo CV; Ruiz DD; Norberto de Souza O; Barros RC Clustering Molecular Dynamics Trajectories for Optimizing Docking Experiments. Comput. Intell. Neurosci 2015, 2015, 916240. [PubMed: 25873944]

(25). Offutt TL; Swift RV; Amaro RE Enhancing Virtual Screening Performance of Protein Kinases with Molecular Dynamics Simulations. J. Chem. Inf. Model 2016, 56, 1923–1935. [PubMed: 27662181]

(26). Motta S; Bonati L Modeling Binding with Large Conformational Changes: Key Points in Ensemble-Docking Approaches. J. Chem. Inf. Model 2017, 57, 1563–1578. [PubMed: 28616990]

(27). Bietz S; Rarey M SIENA: Efficient Compilation of Selective Protein Binding Site Ensembles. J. Chem. Inf. Model 2016, 56, 248–259. [PubMed: 26759067]

(28). Rueda M; Bottegoni G; Abagyan R Recipes for the Selection of Experimental Protein Conformations for Virtual Screening. J. Chem. Inf. Model 2010, 50, 186–193. [PubMed: 20000587]

(29). Rueda M; Totrov M; Abagyan R ALiBERO: Evolving a Team of Complementary Pocket Conformations Rather Than a Single Leader. J. Chem. Inf. Model 2012, 52, 2705–2714. [PubMed: 22947092]

(30). Ben Nasr N; Guillemain H; Lagarde N; Zagury JF; Montes M Multiple Structures for Virtual Ligand Screening: Defining Binding Site Properties-Based Criteria to Optimize the Selection of the Query. J. Chem. Inf. Model 2013, 53, 293–311. [PubMed: 23312043]

(31). Xu M; Lill MA Utilizing Experimental Data for Reducing Ensemble Size in Flexible-Protein Docking. J. Chem. Inf. Model 2012, 52, 187–198. [PubMed: 22146074]

(32). 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, 2664–2679. [PubMed: 25233367]

(33). Huang Z; Wong CF Inexpensive Method for Selecting Receptor Structures for Virtual Screening. J. Chem. Inf. Model 2016, 56, 21–34. [PubMed: 26651874]

(34). Craig IR; Pfleger C; Gohlke H; Essex JW; Spiegel K Pocket-Space Maps to Identify Novel Binding-Site Conformations in Proteins. J. Chem. Inf. Model 2011, 51, 2666–2679. [PubMed: 21910474]

(35). Shao J; Tanner SW; Thompson N; Cheatham TE Clustering Molecular Dynamics Trajectories: 1. Characterizing the Performance of Different Clustering Algorithms. J. Chem. Theory Comput 2007, 3, 2312–2334. [PubMed: 26636222]

(36). Xu S; Zou S; Wang L A Geometric Clustering Algorithm with Applications to Structural Data. J. Comp. Biol 2015, 22, 436–450.

(37). Kinney JJ A Probability and Statistics Companion; John Wiley & Sons, 2009; Chapter 14, pp 211–222.

(38). Dolinsky TJ; Nielsen JE; McCammon JA; Baker NA PDB2PQR: An Automated Pipeline for the Setup of Poisson-Boltzmann Electrostatics Calculations. Nucleic Acids Res 2004, 32, 665–667.

(39). Case DA; Babin V; Berryman JT; Betz RM; Cai Q; Cerutti DS; Cheatham TE, III; Darden TA; Duke R. E; Gohlke H; Goetz AW; Gusarov S; Homeyer N; Janowski P; Kaus J; Kolossva´ry I; Kovalenko A; Lee TS; LeGrand S; Luchko T; Luo R; Madej B; Merz KM; Paesani F; Roe DR; Roitberg A; Sagui C; Salomon-Ferrer R; Seabra G; Simmerling CL; Smith W; Swails J; Walker RC; Wang J; Wolf RM; Wu X; Kollman P AMBER A 2014.

(40). Ponder JW; Case DA Force Fields for Protein Simulations. Adv. Protein Chem 2003, 66, 27–85. [PubMed: 14631816]

(41). Onufriev A; Bashford D; Case DA Exploring Protein Native States and Large-Scale Conformational Changes With a Modified Generalized Born Model. Proteins: Struct, Funct., Bioinf 2004, 55, 383–394.

(42). Eastman P; Pande VS OpenMM: A Hardware-Independent Framework for Molecular Simulations. Comput. Sci. Eng 2010, 12, 34–39.

(43). Weininger D SMILES, a Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules. J. Chem. Inf. Comput. Sci 1988, 28, 31–36.

(44). Chen X; Liu M; Gilson MK BindingDB: a Web-Accessible Molecular Recognition Database. Comb. Chem. High Throughput Screen 2001, 4, 719–725. [PubMed: 11812264]

(45). Vainio MJ; Johnson MS Generating Conformer Ensembles Using a Multiobjective Genetic Algorithm. J. Chem. Inf. Model 2007, 47, 2462–2474. [PubMed: 17892278]

(46). Pettersen EF; Goddard TD; Huang CC; Couch GS; Greenblatt DM; Meng EC; Ferrin TE UCSF Chimera - A Visualization System for Exploratory Research and Analysis. J. Comput. Chem 2004, 25, 1605–1612. [PubMed: 15264254]

(47). Jakalian A; Bush BL; Jack DB; Bayly CI Fast, Efficient Generation of High-Quality Atomic Charges. AM1-BCC Model: I. Method. J. Comput. Chem 1999, 21, 132–146.

(48). Jakalian A; Jack DB; Bayly CI Fast, Efficient Generation of High-Quality Atomic Charges. AM1-BCC Model: Ii. Parameterization and Validation. J. Comput. Chem 2002, 23, 1623–1641. [PubMed: 12395429]

(49). Wang J; Wolf RM; Caldwell JW; Kollman PA; Case DA Development and Testing of a General Amber Force Field. J. Comput. Chem 2004, 25, 1157–1174. [PubMed: 15116359]

(50). The UniProt Consortium. UniProt: the Universal Protein Knowledgebase. Nucleic Acids Res 2017, 45, D158–D169. [PubMed: 27899622]

(51). Eswar N; Webb B; Marti-??Renom MA; Madhusudhan MS; Eramian D; Shen M; Pieper U; Sali A Comparative Protein Structure Modeling Using MOD-ELLER. Curr. Protoc. Protein. Sci 2007, 50, 2.9.1–2.9.31.

(52). Bakan A; Meireles LM; Bahar I ProDy: Protein Dynamics Inferred From Theory and Experiments. Bioinformatics 2011, 27, 1575–1577. [PubMed: 21471012]

(53). Lang P; Brozell SR; Mukherjee S; Pettersen E; Meng EC; Thomas V; Rizzo RC; Case DA; James T; Kuntz ID DOCK 6: Combining Techniques to Model RNA-Small Molecule Complexes. RNA 2009, 15, 1219–1230. [PubMed: 19369428]

(54). Durrant JD; Votapka L; Sørensen J; Amaro RE POVME 2.0: An Enhanced Tool for Determining Pocket Shape and Volume Characteristics. J. Chem. Theory Comput 2014, 10, 5047–5056. [PubMed: 25400521]

(55). O'Donoghue P; Luthey-Schulten Z Evolutionary Profiles Derived From the QR Factorization of Multiple Structural Alignments Gives an Economy of Information. J. Mol. Biol 2005, 346, 875–894. [PubMed: 15713469]

(56). Humphrey W; Dalke A; Schulten K VMD - Visual Molecular Dynamics. J. Mol. Graphics 1996, 14, 33–38.

(57). van der Walt S; Colbert SC; Varoquaux G The NumPy Array: A Structure for Efficient Numerical Computation. Comput. Sci. Eng 2011, 13, 22–30.

(58). Bajorath J; Kitson DH; Hagler AT; Kraut J The Electrostatic Potential of Escherichia Coli Dihydrofolate Reductase. Proteins: Struct., Funct., Genet 1991, 11, 1–12. [PubMed: 1961697]

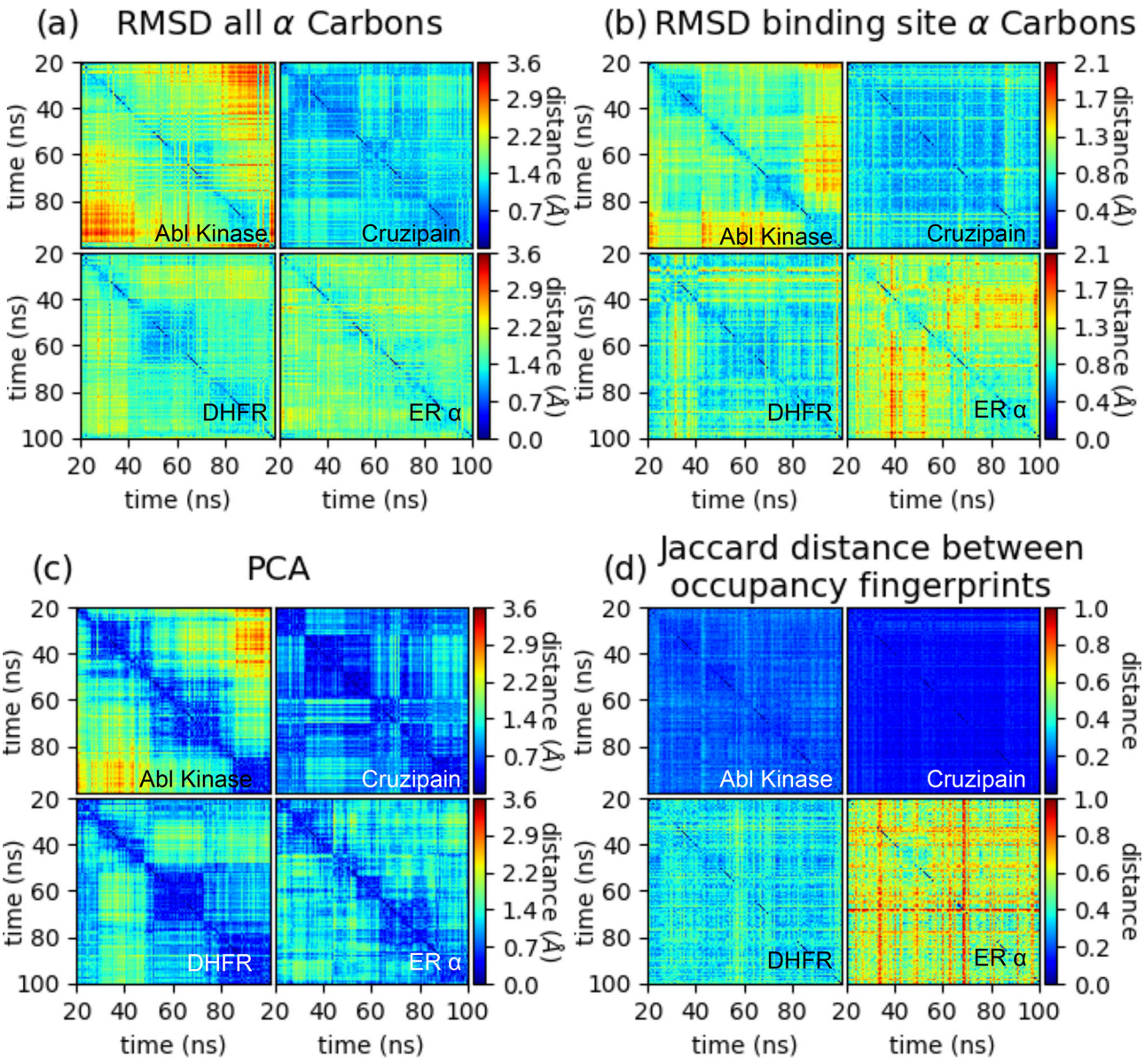


Figure 1:

Distance matrices for MD simulations of Abl kinase (top left), Cruzipain (top right), DHFR (bottom left), and ER α (bottom right) based on (a) the RMSD between α carbons in the whole protein, (b) the RMSD between α carbons in the binding site, (c) PCA, (d) the Jaccard distance between occupancy fingerprints.

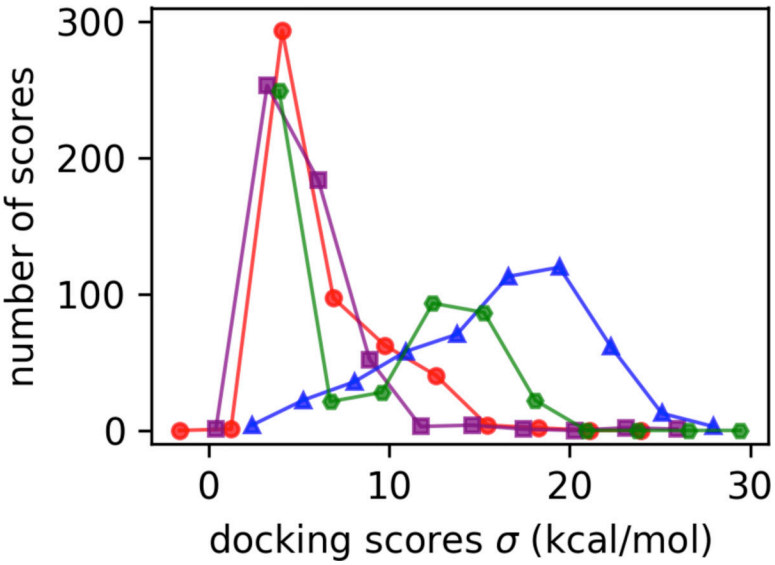


Figure 2: The standard deviation of docking scores.

Histograms of the standard deviation, σ, of docking scores for different ligands binding to the population of snapshots for each protein: Abl kinase (red circles), cruzipain (purple square), DHFR (blue upwards triangles), ER α (green hexagons). The standard deviation of docking scores is computed for each ligand. A histogram of these standard deviations is shown.

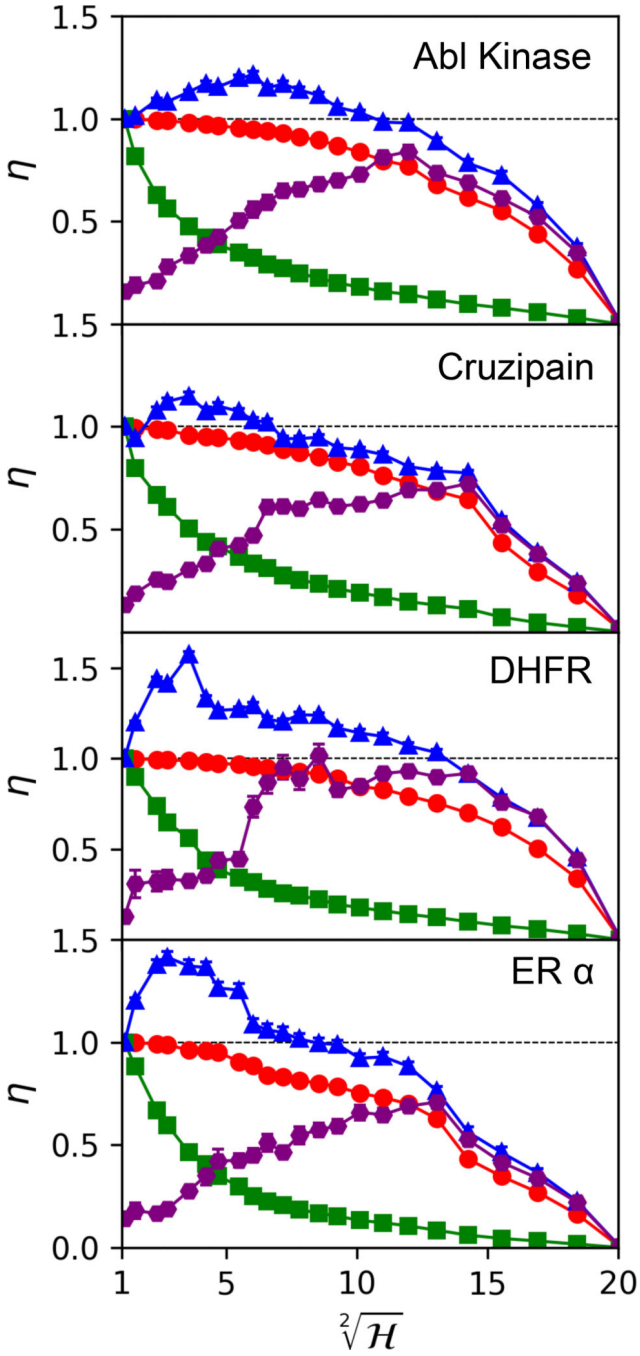


Figure 3:

Comparing allocation schemes for computing ensemble averages: archetypal senatorial (purple hexagons), optimal (green squares), proportional (red circles), and senatorial (blue upwards triangles). Hierarchical clustering was performed based on the Jaccard distance between occupancy fingerprints and the complete linkage criterion.

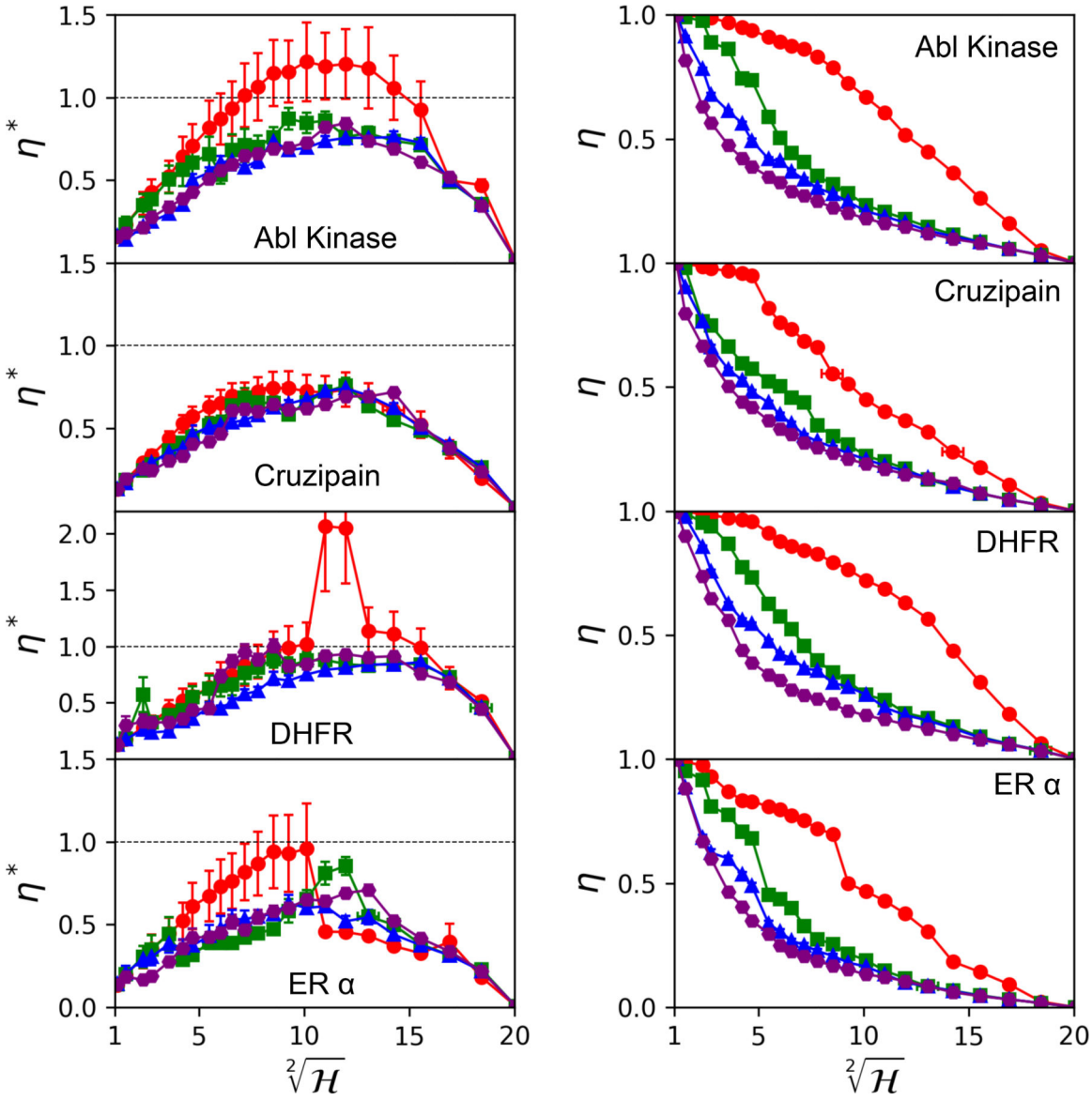


Figure 4: Comparing linkage algorithms for computing ensemble averages based on archetypal senatorial (left) and optimal (right) allocation.

Clustering was performed based on the Jaccard distance between occupancy fingerprints and complete (purple hexagons), weighted (blue upwards triangles), average (green squares), and single (red circles) linkage algorithms.

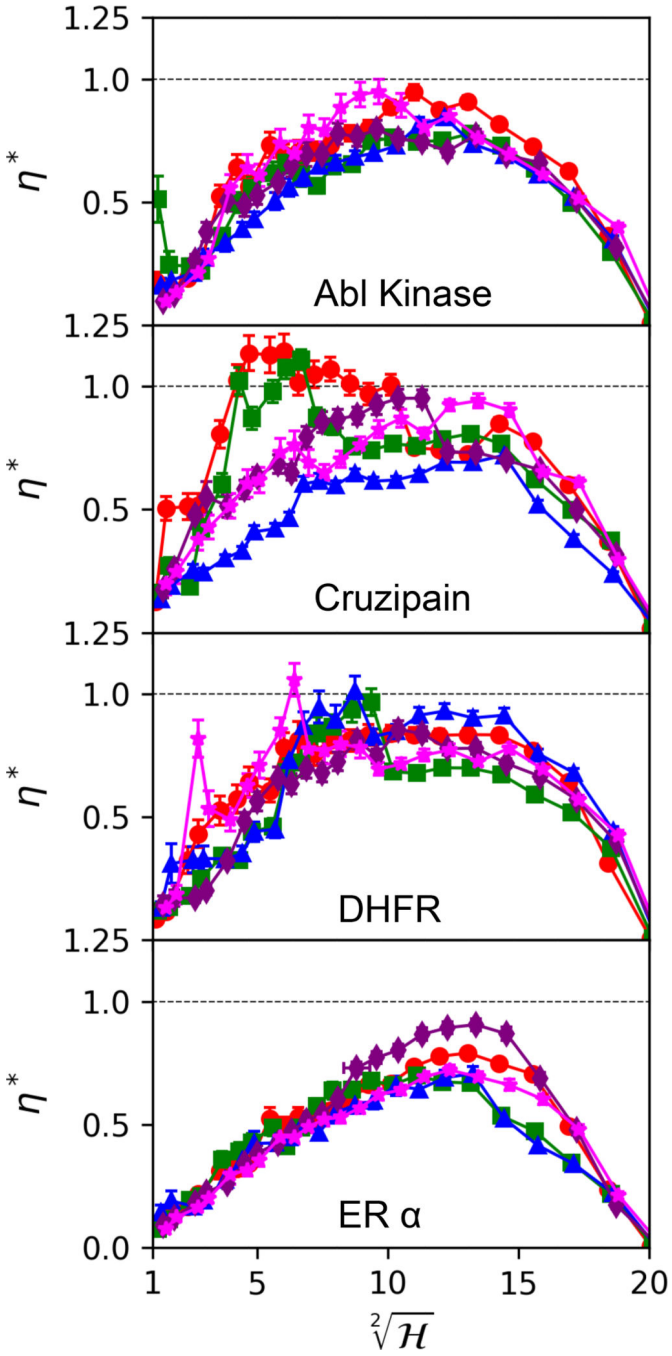


Figure 5:

Comparing types of distances for computing ensemble averages using archetypal senatorial allocation: the RMSD between all α carbons (red circles), binding site α carbons (purple hexagons), and binding site heavy atoms (magenta downwards triangles); the Jaccard distance between overlap fingerprints (blue upwards triangles); and PCA (green squares).

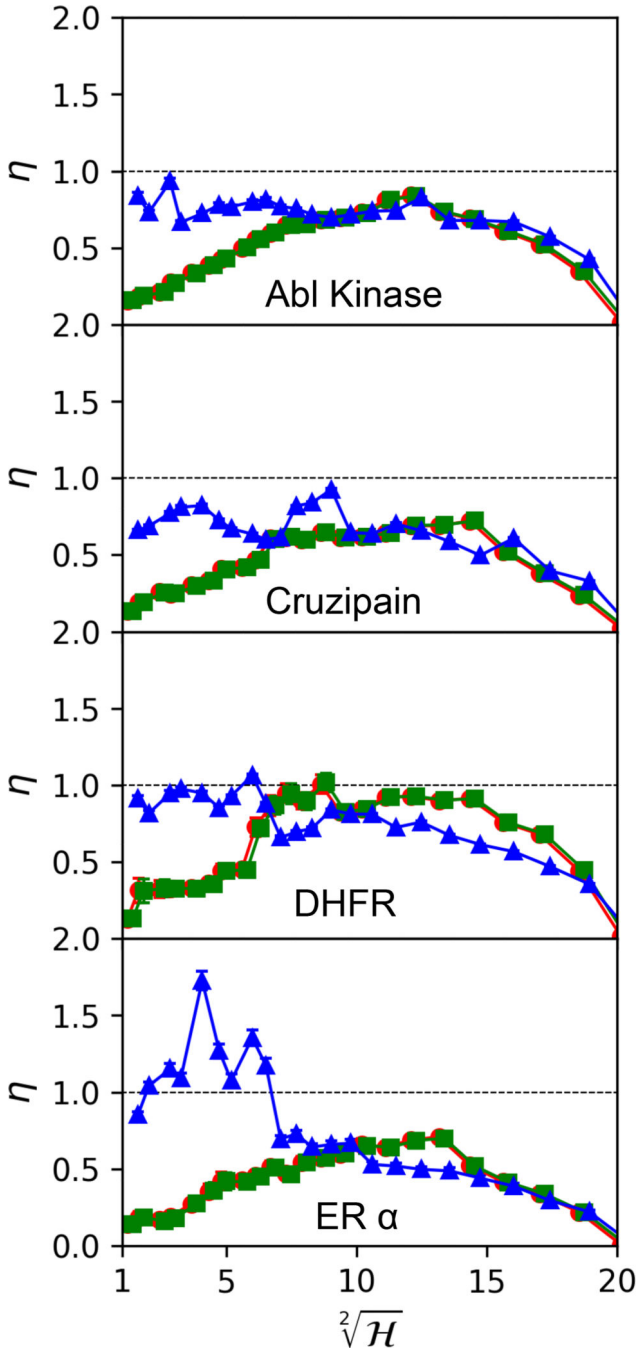


Figure 6:

Comparing the stratification efficiency of archetypal senatorial allocation for different summary statistics: the ensemble average (red circles), exponential average (green squares), and minima (blue upwards triangles). Clustering was performed based on the Jaccard distance between occupancy fingerprints and the complete linkage criterion.

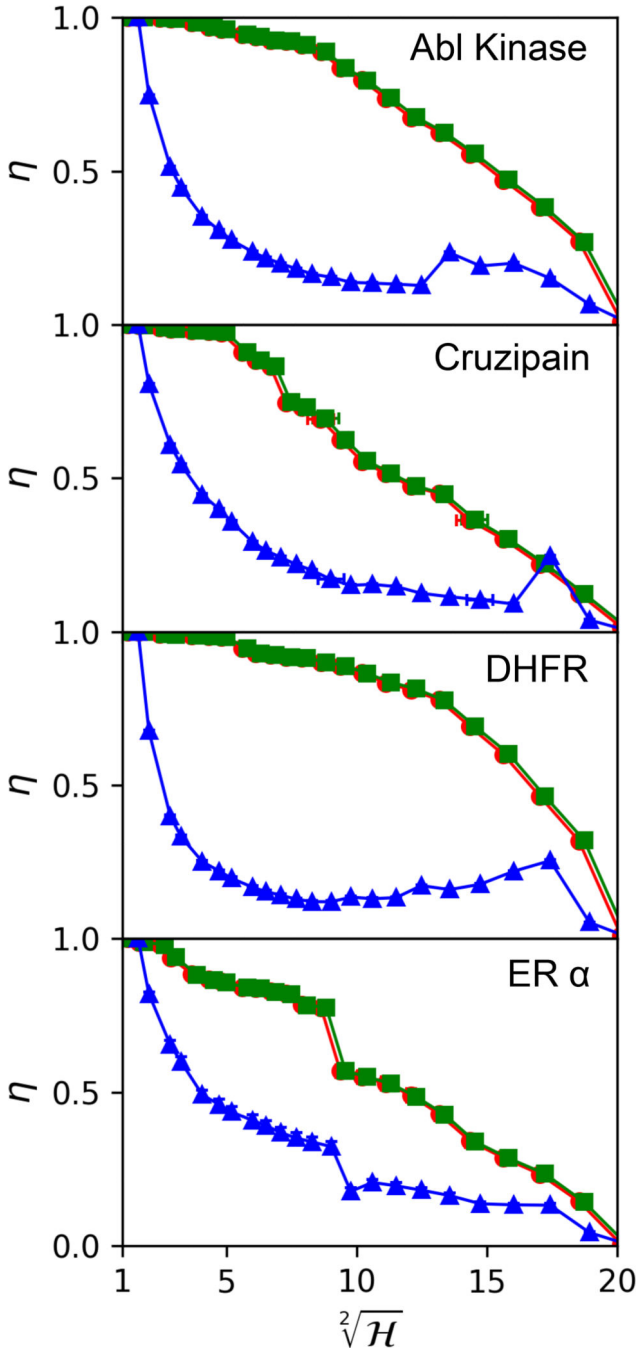


Figure 7:

Comparing the stratification efficiency of proportional allocation for different summary statistics: the ensemble average (red circles), exponential average (green squares), and minima (blue upwards triangles). Clustering was performed based on the Jaccard distance between occupancy fingerprints and the single linkage criterion.

Table 1:

Protein information.

Protein system	UniProt ID	Minimum sequence identity	Reference chain
Abl Kinase	P00520	100	1OPJ (A)
Cruzipain	P25779	90	1ME4 (A)
DHFR	P13922	90	1J3J (A)
ERα	P03372	99	1X7E (A)