JCIM JOURNAL OF CHEMICAL INFORMATION AND MODELING

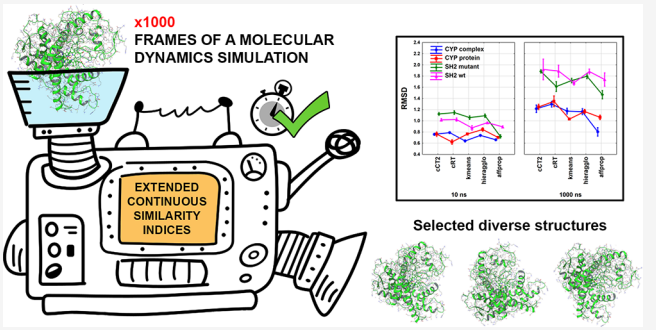
pubs.acs.org/jcim Article

Molecular Dynamics Simulations and Diversity Selection by Extended Continuous Similarity Indices

Anita Rácz,[‖] Levente M. Mihalovits,[‖] Dávid Bajusz, Károly Héberger,* and Ramón Alain Miranda-Quintana*

Cite This: J. Chem. Inf. Model. 2022, 62, 3415−3425 Read Online

ABSTRACT: Molecular dynamics (MD) is a core methodology of molecular modeling and computational design for the study of the dynamics and temporal evolution of molecular systems. MD simulations have particularly benefited from the rapid increase of computational power that has characterized the past decades of computational chemical research, being the first method to be successfully migrated to the GPU infrastructure. While new-generation MD software is capable of delivering simulations on an ever-increasing scale, relatively less effort is invested in developing postprocessing methods that can keep up with the quickly expanding volumes of data that are being generated. Here, we introduce a new idea for sampling frames from large MD trajectories, based on the recently introduced framework of extended similarity indices. Our approach presents a new, linearly scaling alternative to the traditional approach of applying a clustering algorithm that usually scales as a quadratic function of the number of frames. When showcasing its usage on case studies with different system sizes and simulation lengths, we have registered speedups of up to 2 orders of magnitude, as compared to traditional clustering algorithms. The conformational diversity of the selected frames is also noticeably higher, which is a further advantage for certain applications, such as the selection of structural ensembles for ligand docking. The method is available open-source at https://github.com/ramirandaq/MultipleComparisons.



1. INTRODUCTION

With the exponential increase in computer hardware capacity, the application of molecular dynamics (MD) has become an essential tool in computational chemistry and related studies. The usage of graphical processing units (GPUs) extends the feasible length of MD simulations,[1] allowing researchers to simulate processes even in the microsecond time scale.[2] While many disciplines benefit from MD simulations, such as medicinal chemistry,[3] materials science,[4] biophysics,[5] or biochemistry,[6] this paper focuses mainly on the first one. MD is commonly used to examine specific events and properties on a molecular basis, most notably structural changes,[7] structural stability,[8] chemical reactions,[9] and dynamics of atomic-level phenomena.[4] Coupled with statistical thermodynamics, MD simulations are able to account for energies of simulation-related processes, as well.[10] Additionally, structures obtained from MD trajectories aid other computational methods, too. Protein structures extracted from trajectories help to overcome the limitations of rigid ligand docking,[11] enabling the ligands to fit into multiple protein structures.[12] This approach is commonly termed ensemble docking[13] and was shown to increase the performance of structure-based virtual screening, a popular method for early hit discovery in rational drug design.[14] Moreover, the stability of these protein−ligand complexes can be verified also by MD simulations.[15] Besides the structural data, the output trajectories of the simulations require postprocessing methods to extract valuable results. Common procedures are simulation event analysis, simulation quality analysis, and trajectory clustering.[16,17] The latter is used frequently to obtain representative structures of the given trajectory; however, frame selection and clustering are highly nontrivial tasks, which can be heavily problem-dependent. The main questions are which indices should be used for the selection (the most frequent choice is the root mean squared distance, RMSD) and how the selection should be made (most different structures or most common structures). Commercial MD software packages (AMBERtools, Desmond, NAMD) usually contain built-in clustering programs;[16] however, their performance is hard to measure, and their applicability for different trajectory formats is ponderous.

Received: April 15, 2022
Published: July 14, 2022

https://doi.org/10.1021/acs.jcim.2c00433
J. Chem. Inf. Model. 2022, 62, 3415−3425

Therefore, as an alternative to clustering methods, we have developed a diversity picker based on the recently introduced extended continuous similarity indices, which requires only the coordinates of the atoms in the extracted snapshots of the trajectory and can be implemented easily. The algorithm is inspired by the diversity pickers commonly applied in cheminformatics to sample large chemical spaces, usually based on the use of binary molecular fingerprints.[18] The various versions of the extended similarity indices[18−20] have shown great promise in the problems of diversity selection[21] and exploration of large and various data sets[22,23] including complex biological ensembles.[24] The keys to this success are the ability of the extended indices to quantify similarities between any number of objects and the fact that they can do so with *linear* scaling.

In ref 24, we explored the application of extended similarity indices to the *classification* of conformations in biological ensembles. To this end, we developed a novel hierarchical agglomerative clustering algorithm that successfully distinguished between conformations corresponding to different stages along multiple folding pathways. However, the fact that we had to start from a clustering step means that this approach scaled as $O(N^2)$. Moreover, we only considered extended similarity indices defined over binary vectors.[24] That is, there was the need to perform a preprocessing step transforming the real-valued coordinates into bit-vectors via contact maps. This is problematic for three reasons: a) this preprocessing step can be time-consuming, b) it is not clear *a priori* which residues should be selected to provide an optimum contact map representation, and c) there is an intrinsic information loss when we go from real-valued to binary quantities. In this work, we overcome this latter deficiency by defining extended continuous similarity indices. Hence, a novel extended similarity-based algorithm was developed to efficiently select diverse and representative structures from long MD simulations. We evaluated the new method on case studies of MD simulations with different lengths and system sizes. The obtained trajectories were evaluated, and its performance was compared with common clustering algorithms as benchmarks. Notably, the developed algorithm was used for the postprocessing of a 100 μs long MD simulation of the SARS-CoV-2 main protease (PDB: 6Y84) to demonstrate the potential benefits of the extended continuous similarity indices, including their excellent scalability. The latter is especially relevant today, as the increase in computational capacities and access to powerful supercomputers enable access to unprecedented simulation times, but the efficiency of postprocessing methods rarely matches the capabilities of the core simulation programs. As the emphasis of this new method is on *sampling*, instead of classifying, the different conformations do not require any clustering step, which makes it very attractive, as the overall approach scales as $O(N)$. Furthermore, in this manuscript, we circumvent all these issues by using a generalization of the extended similarity indices that is suitable for real-valued quantities. This means that the only preprocessing required is a simple normalization of the coordinates, that we can include as many atomic coordinates as possible, and that we are not losing any information while performing the sampling. Our method is available open-source at https://github.com/ramirandaq/MultipleComparisons.

2. MATERIALS AND METHODS

2.1. Extended Continuous Similarity for MD Simulation Data. The analysis of MD data differs from the more conventional cheminformatic applications of our extended similarity indices due to the fundamentally different nature of the input data in both cases. In fact, we regard the recently introduced extended continuous similarity indices as a set of new similarity measures altogether, since they include completely original concepts to allow for the similarity calculation of an arbitrary number of continuous vectors.[20] Nonetheless, we decided to keep the names of the existing similarity metrics that served as their basis (*e.g.*, Russell-Rao, Jaccard-Tanimoto, *etc.*), so that they can be more easily traced back to the widely known, "traditional" measures. For reference, the original extended (or n-ary) similarity indices were defined over dichotomous variables[18,21] (*e.g.*, binary molecular fingerprints), which simplified the connection to the standard binary indices, such as the Tanimoto coefficient. However, since we are now dealing with atomic coordinates, we need to be able to efficiently process real-valued vectors. This also means that our similarity indices are capable of processing these vectors in other fields as well.[20] The real-valued vectors in this situation demand two key aspects: first, we need to find a suitable way to normalize the coordinate values, since the extended continuous indices are defined over the $[0, 1]$ interval. There are many (in principle, infinite) ways to perform a normalization, but the nature of this problem leads to a very natural decision. As noted above, the ultimate goal of our approach is to select different conformations that are, at the same time, as representative and diverse as possible. We also want our selection to be in accordance with standard approaches used to assess the quality of the frames selected, which is typically based on the root-mean-square distance (RMSD) values of the chosen structures. In other words, our normalization scheme must be consistent with the RMSD calculation, so the scaling procedure should not interfere with the selection algorithm. The notion of consistency[25−27] is central to the work with similarity indices and in this particular case can be expressed quite simply: let $\{q_i^{(a)}\}$ and $\{q_i^{(b)}\}$ be the coordinates corresponding to any two frames, a and b, then the normalization function $n(q_i)$ must satisfy

$$\left[\sum_i (q_i^{(a)} - q_i^{(b)})^2\right]\left[\sum_i (n(q_i^{(a)}) - n(q_i^{(b)}))^2\right] \geq 0 \tag{1}$$

Notice that any function that satisfies this inequality will preserve the intrinsic ordering of the conformations, so we just need to find a suitable functional form. Luckily, this can be easily done by taking

$$n(q_i) = \frac{q_i - \min\{q_i\}}{\max\{q_i\} - \min\{q_i\}} \tag{2}$$

It is critical that the minimum and maximum in the normalization function are taken over all the coordinates of all the conformations, which is the only way to guarantee a proper uniform scaling.

The second key step in the definition of the real-valued extended indices is how to obtain the analogues of the 1-similarity, 0-similarity, and dissimilarity counters introduced for the dichotomous n-ary indices.[18,21] There are several variants that could establish a one-to-one correspondence between the binary and the continuous case, but once again the nature of the problem at hand suggests a simple solution. In our recent work that applied the original extended similarity indices to the study of the conformational landscape of several biomolecules,[24] it was shown that a simple column-wise sum of the matrix of

https://doi.org/10.1021/acs.jcim.2c00433
J. Chem. Inf. Model. 2022, 62, 3415−3425

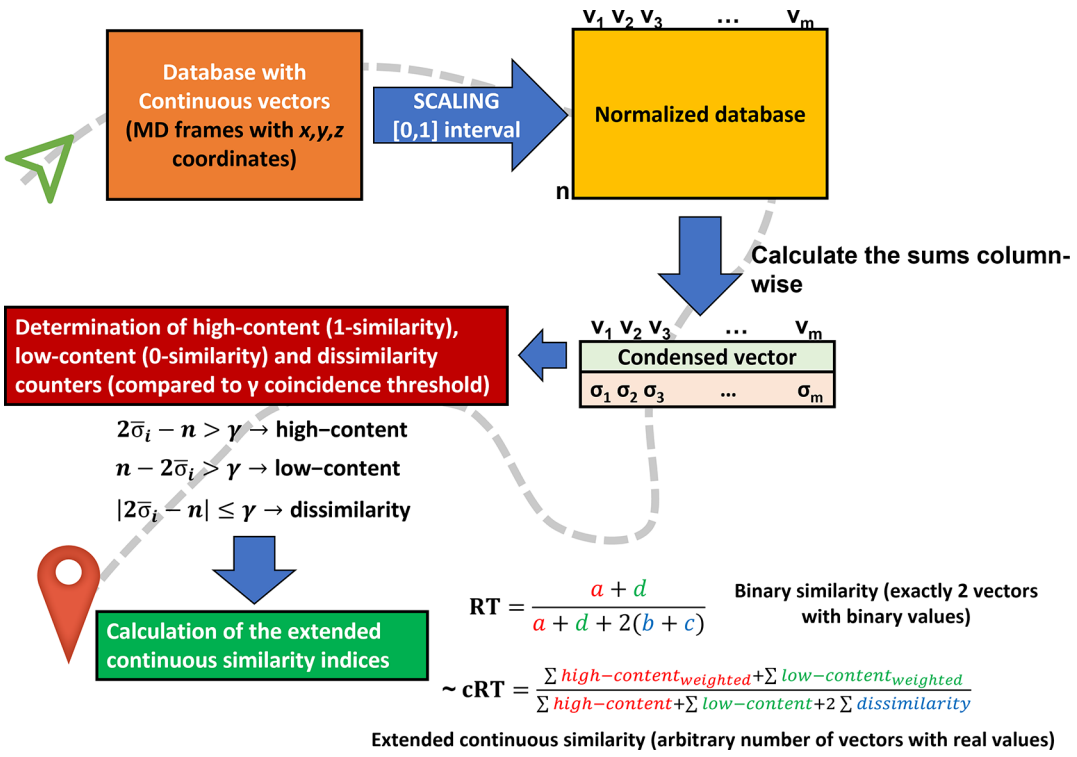


Figure 1. Calculation steps of the extended continuous similarity indices from a database with continuous vectors.

conformations was the best alternative. Hence, this will be the procedure that we will follow here, with the sum of each column of the normalized matrix taken as the central elements of the formalism.[20]

The first step is then to store the coordinates of each conformation in a row vector. Next, we arrange all these vectors in a matrix, where the rows are conformations, and the columns correspond to the coordinates of each atom. Then we proceed to normalize the entries in this matrix using eq 2 and generate a row vector containing the sum of each column of the normalized matrix. This is the key input required to calculate the extended similarity of the set of conformations. With this, we can proceed to classify the entries of the column sum vector in high-content similarity, low-content similarity, or dissimilarity counters. Finally, after appropriately weighting these counters, we can calculate any desired n-ary similarity index.

Figure 1 illustrates the most important steps from the starting point to the generation of extended continuous similarity indices. The details of each step can be followed in the Supporting Information with an example calculation of the nonweighted extended continuous Rogers-Tanimoto index.

2.2. Diversity Selection with ECS-MeDiv. The diversity selection algorithm proposed here is rooted in two central ideas previously explored in unrelated applications of the extended similarity indices: (i) medoid determination and (ii) selecting the most diverse structures from a given data set. It has been shown that the extended similarity indices provide a very attractive solution to the problem of finding the most representative element of a set (*e.g.*, the medoid).[22,24] This is done by calculating the complementary similarity of each element (that is, the extended similarity of the original set minus the corresponding element) and picking the point with lowest complementary value. The complementary similarity is a measure of how "connected" an element is to all others in the set. A "central" element (*e.g.*, the medoid) will be heavily connected to the remaining elements; hence, if we remove it and calculate the similarity of the other points, this value will be low (compared to the removal of other elements). Consequently, a low complementary similarity is synonymous with the central character of an element in the set. This naturally leads to a ranking of elements, from more representative or "central" to less representative or "outliers".

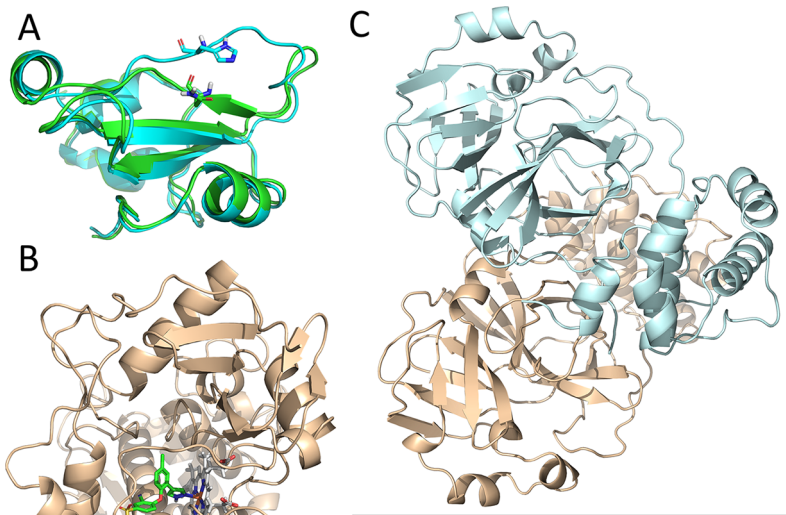
Notice, however, that this ranking is not sufficient for our present purposes. Selecting only central conformations is a bad sampling strategy because it will heavily favor native-like structures, which correspond to a very narrow region of the conformational space. Meanwhile, sampling just from the outlier structures means that we are going to miss important low-energy structures that are close to the native state. These problems cannot be solved by just using a diversity picker because we would still need to solve the problem of how to initialize the algorithm (which structure should we pick first?). That specific selection must be made reasonably and consistently, otherwise it can lead to nonreproducible errors and oversampling of outlier regions of the conformational space.

We approached these problems by combining these approaches in our *Extended Continuous Similarity − Medoid Diversity* (ECS-MeDiv) algorithm. The general strategy is as follows:

1. Select the medoid of the set as the starting conformation for the diversity picker.
2. Repeatedly, given the set of conformations already picked $C = \{c_1,...,c_n\}$ select the conformation c' such that the set $\{c_1,...,c_n,c'\}$ has the lowest possible extended similarity.

2.1 If there are several conformations $c',c'',...$ that lead to the same extended similarity when added to the preselected conformations, pick the one that has the lowest value of the average binary similarity with all the elements of the preselected set.

https://doi.org/10.1021/acs.jcim.2c00433
J. Chem. Inf. Model. 2022, 62, 3415−3425



	SH2	CYP	3CLPro
# amino acids	97	469	644
# atoms (protein)	1561 (*wt*) 1564 (*N642H*)	7643	9364
# atoms (total)	17127 (*wt*) 17130 (*N642H*)	50565	158788
Sim. time (ns)	10 / 1000	10 / 1000	100000
# frames	1000	1000	100000

Figure 2. Model systems: **A**) Wild-type (green, PDB: 6MBW) and N642H mutant (cyan, PDB: 6MBZ) structure of the SH2 domain of STAT5B, with residue 642 highlighted as sticks. In the mutant structure, the uppermost β sheet is disconnected because of the mutation.[31] **B**) CYP 2C9 in complex with a small-molecule inhibitor (green sticks, PDB: 5K7K).[37] This system was simulated in the holo (liganded, "CYP complex") and apo (unliganded, "CYP protein") states, and the heme cofactor (white sticks) was kept in both simulations. **C**) Dimer structure of the SARS-CoV-2 main protease in the apo (unliganded) state (PDB: 6Y84). The monomers are shown in different colors. The table contains the number of amino acids, atoms, and frames for the MD simulations of each model system.

3. Continue until the desired number of conformations has been selected (or the conformation pool has been exhausted).

This algorithm improves upon our Max_nDis picker[21] because step 2.1 allows for a more thorough selection of the diverse conformations by effectively serving as a tiebreaker between conformations with the same extended similarity with respect to the preselected set. That is, we still use the minimization of the extended similarity (step 2) as the driving force of the algorithm, but step 2.1 adds an extra layer that leads to an even more diverse set in the end. The focus on the extended similarity also means that we do not need to generate the binary distance matrix, hence guaranteeing an efficient global exploration of all the conformations. Having to calculate a small (if any) number of binary similarities at each step implies that the ECS-MeDiv algorithm scales linearly for the selection in cases like the ones considered here, where one is only interested in a comparatively small number of conformations (5−10 out of 1000 or 100000). Notice also how by starting from the medoid (also a linearly scaling step) we ensure a uniform sampling of the MD trajectories.

2.3. Data Sets. To showcase the utility of our method, we have performed equilibrium MD simulations on two model systems (case studies) of medicinal chemistry relevance. The first system was the SH2 domain of the STAT5B transcription factor. SH2 domains are small, modular protein units that recognize phosphotyrosine-containing peptide motifs in a highly selective manner and are widely utilized in cellular signal transduction.[28] STATs (Signal Transducers and Activators of Transcription) are a small family of multidomain proteins with pivotal roles in the regulation of DNA transcription.[29] Dimer formation via the SH2 domain is a primary requirement of STAT function, and the inhibition of this protein−protein interaction was identified as a point of pharmaceutical intervention for several oncological indications, such as acute myeloid leukemia.[30] This is especially relevant when STATs themselves act as oncogenesis drivers, upon point mutations in the SH2 domain.[29] One such driver mutation of STAT5B, namely N642H, was recently identified to induce a significant conformational change in the SH2 domain.[31] Our first model system involved the simulation of the wild-type and N642H mutant SH2 domains of STAT5B, representing two variants of a small and relatively flexible protein (Figure 2A).

The second model system is a structure of the 2C9 isoenzyme of the cytochrome P450 (CYP) protein family of heme-thiolate enzymes.[32] CYPs have key roles in the metabolic processes of virtually all organisms by catalyzing a vast range of reactions, including the activation of molecular oxygen. In humans, hepatic CYP enzymes are the main drivers of drug metabolism,[33] with the 2C9 isozyme being responsible for the hepatic clearance of 12−16% of clinically relevant drugs.[34] CYP 2C9 is a highly relevant ADME (Absorption, Distribution, Metabolism, Excretion) target, and significant efforts are dedicated to the *in silico* prediction of the affinity of small molecules to this enzyme.[35,36] Here, we have simulated the dynamics of the CYP 2C9 isozyme (PDB: 5K7K[37]) with and without its cocrystallized small-molecule inhibitor, representing a larger, more rigid system (Figure 2B).

Finally, we have demonstrated the performance of our method on a long simulation of an even larger system. To that

https://doi.org/10.1021/acs.jcim.2c00433
J. Chem. Inf. Model. 2022, 62, 3415−3425

end, we have downloaded the publicly available, 100 μs trajectory of the SARS-CoV-2 main protease performed on the Anton 2 supercomputer[38] by D. E. Shaw Research. The main protease (3CLPro) cleaves the replicated viral polypeptides into functional proteins, and it was identified early as a potential antiviral target (Figure 2C).[39] Since the outbreak of the COVID-19 pandemic, it was in the forefront of drug discovery efforts, from state-of-the-art crystallographic fragment screening approaches by academia[40] to the recent breakthrough of Pfizer to provide the first approved antiviral drug specifically developed against COVID-19.[41] Here, the simulation of the 3CLPro dimer serves as a case study to represent the current state of the art in terms of accessible simulation length on a fairly large molecular system.

2.4. Molecular Dynamics Simulations. Starting structures for the four proteins were extracted from the PDB structure 5K7K (CYP complex and CYP protein),[37] 6MBW (SH2 wild type, wt), and 6MBZ (SH2 N642H mutant).[31] Protein preparation was performed with Schrödinger Maestro's Protein Preparation Wizard.[42] For the wild-type and mutant SH2 domain structures, chains B and A were kept, respectively. The CYP complex structure contains the HEM cofactor and a bound small-molecule ligand 6RJ, while the CYP protein was created from the same PDB file with the deletion of 6RJ. System preparation was carried out with Desmond's system builder. All structures were immersed into a minimized, buffer-sized orthorhombic TIP3P water-box. System neutralization was achieved by chlorine ion addition. The assigned force field was set to OPLS3e.[43] The resultant solvated systems were relaxed by Desmond's relaxation protocol. Molecular dynamics simulations were performed by Desmond in the NVT ensemble at 298.15 K, applying Nosé−Hoover temperature regulation.[44,45] Two unbiased MD simulations were carried out for every system with simulation times of 10 and 1000 ns, respectively. These correspond to scenarios of a quick and more thorough conformational sampling by MD. Every trajectory consists of 1000 frames, which were later used for the similarity examinations and the diversity selection (clustering). The molecular dynamics simulation of the SARS-CoV-2 main protease (100 μs) was performed based on the PDB structure 6Y84 on the Anton 2 supercomputer[38] by D. E. Shaw Research, and the trajectory was obtained from their web site: https://www.deshawresearch.com/downloads/download_trajectory_sarscov2.cgi/.

All the evaluations and diversity selection studies were based on the C, Cα, and N atoms of the protein backbone. The x, y, and z coordinates of these atoms were extracted with an in-house script, using VMD[46] to convert the Desmond trajectory files into a series of mol2 format structures. The script is available at https://github.com/ramirandaq/MultipleComparisons.

2.5. Benchmark Clustering of the MD Simulations. The benchmark clustering was done with (i) the affinity propagation algorithm, as implemented in Schrödinger Maestro (trj_cluster.py),[47] (ii) the hierarchical agglomerative approach,[48] and (iii) the k-means method.[49] The latter two are implemented in the cpptraj module of AMBERtools.[16] The number of clusters was set to 5, 6, 7, 8, 9, and 10, respectively. The representative frames of the resultant clusters with the C, Cα, and N coordinates were written out for the similarity calculations. All the other parameters of the clustering methods were set to their default values.

2.6. Statistical Analysis. The similarity values calculated with 16 different extended continuous similarity indices were given for each system (4) and each simulation length (2). The data set consists of tables with the x, y, and z atomic coordinates in the columns and the actual coordinate values throughout the 1000 frames in the rows in a sequential order. The structural diversity of the MD frames was visualized with the t-distributed stochastic neighbor embedding (t-SNE) method based on the atomic coordinates.[50] This allowed us to detect the differences of the traversed conformational space between the simulations of different lengths.

Analysis of variance (ANOVA) was used for the statistical comparison of the factors such as the (i) similarity indices (16), (ii) simulation lengths (2), or (iii) molecular systems (4). Moreover, we have analyzed the results of the diverse set selections with ANOVA, as well. STATISTICA 13 software (TIBCO) was used for the statistical analysis of the data sets. t-SNE vectors (dimensions) were calculated in the KNIME Analytics Platform,[51] while the figures were created with GNUplot.[52]

2.7. RMSD Calculations. The average pairwise root-mean-square deviation (RMSD) for the given set of selected frames was calculated using an in-house script, following eqs S1 and S2. Note that RMSD refers to that average pairwise RMSD value throughout the publication. The standard deviation (std) was computed based on the difference between the average pairwise RMSD and the specific pairwise RMSDs (see eq S3).

3. RESULTS AND DISCUSSION

To explore the applicability of the proposed methodology, we examined four different biologically relevant systems (further on referred to as the CYP complex, CYP protein, SH2 wild-type, and SH2 mutant, *cf.* section 2.2, Figure 2) with 10 and 1000 ns long MD simulation lengths. The coordinates of the backbone atoms (C, Cα, and N) were extracted for the 1000 frames of each simulation, and the extended continuous similarity indices were calculated for each system and simulation length. Figure 3 shows the complete workflow of the study with the most relevant steps.

Sixteen different similarity indices were calculated for each MD run and evaluated with ANOVA. In the next step, we applied our diversity picker on the MD simulations for each case study and compared it to benchmark clustering algorithms based on the RMSD values. Finally, we showed how the extended similarity-based diversity selection protocol works on an

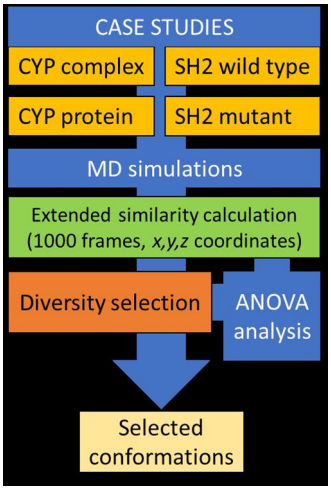


Figure 3. Major steps of our workflow to examine and compare the usage of various continuous similarity indices.

https://doi.org/10.1021/acs.jcim.2c00433
J. Chem. Inf. Model. 2022, 62, 3415−3425

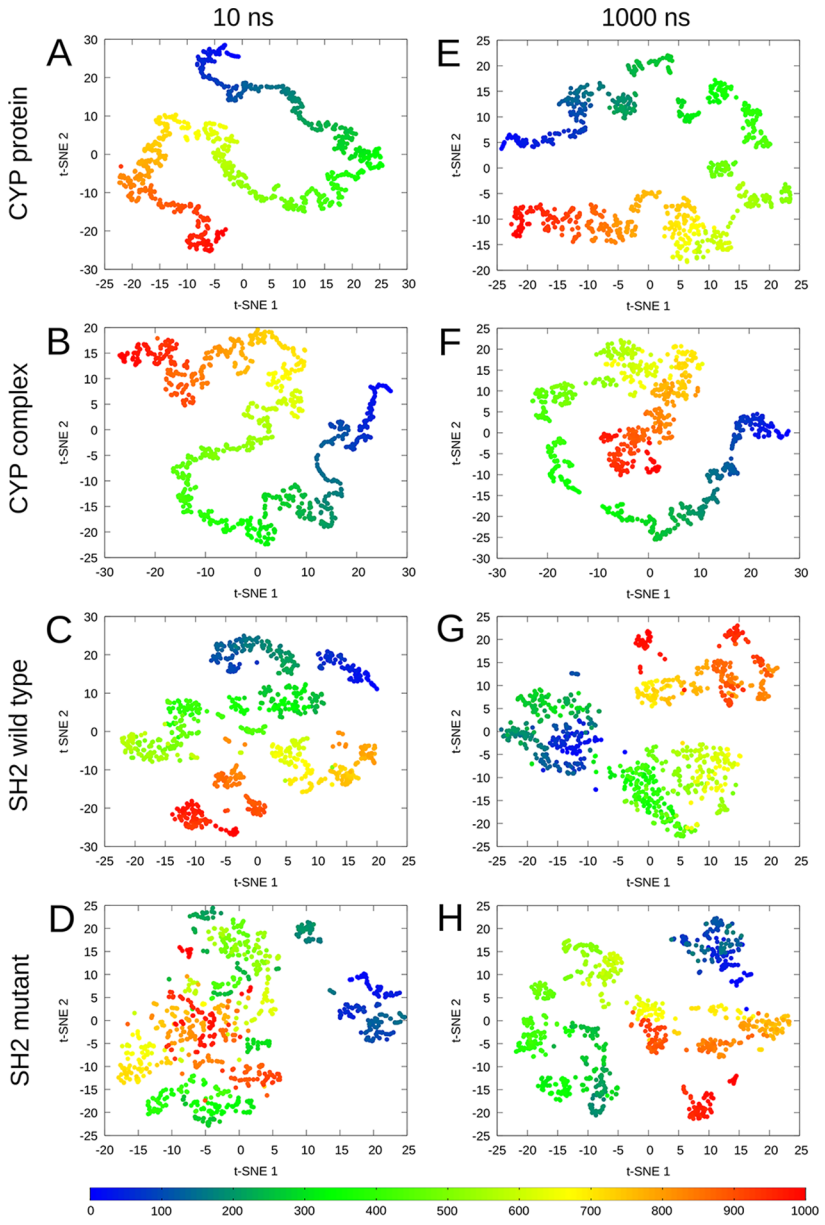


Figure 4. t-SNE plots of the 10 and 1000 ns long MD simulations' coordinates of the CYP 2C9 protein (A, E), CYP 2C9-ligand complex (B, F), wild-type (C, G), and N642H mutant SH2 domains (D, H). The data points are gradually colored from blue (first frame) to red (1000th frame) following the progression of simulation time.

exceptionally long MD trajectory of the main protease (3CLPro) of SARS-CoV-2 with 100,000 frames.

3.1. Evaluation of the MD Simulation Results. As a basic description of the MD simulations, we have visualized the trajectories on t-SNE plots, where the 1000 snapshots of the MD simulations, described by the *x*, *y*, and *z* coordinates, are plotted as data points on a two-dimensional graph (Figure 4). The data points are colored gradually as the simulation time progresses. This representation visualizes the sampled conformational space and gives valuable information on the heterogeneity of the MD frames during the simulation.

A simulation length of 10 ns is considered fairly short, and only slight conformational changes are expected during this period of time. However, it is common to use comparable simulation lengths for quick, equilibrium conformational sampling. Comparing the 10 ns long MDs across the different systems (CYP and SH2), the main differences arise from the size of the proteins. The SH2 domain consists of 97 residues and is much more flexible than the CYP systems with around 500 residues and an overall more rigid structure. The 10 ns t-SNE plots of the CYP systems reveal linearly progressing trajectories, in which the conformational space is discovered more sparsely than in the case of the SH2 proteins, where the frames are more scattered, especially in the case of the N642H mutant.

The 1000 ns simulation length is more suitable for examining protein dynamics, and overall, more significant conformational changes are expected in this case. The t-SNE plots of the 1000 ns simulations show similar trends as the 10 ns ones, with somewhat poorer sampling and more diffuse but still linear trajectories for the CYP systems. The degree of scattering is comparable between the ligand-bound and ligand-free systems, so the bound ligand does not seem to constrain the dynamics of the protein. Interestingly, the t-SNE plot of the wild-type SH2

https://doi.org/10.1021/acs.jcim.2c00433
J. Chem. Inf. Model. 2022, 62, 3415−3425

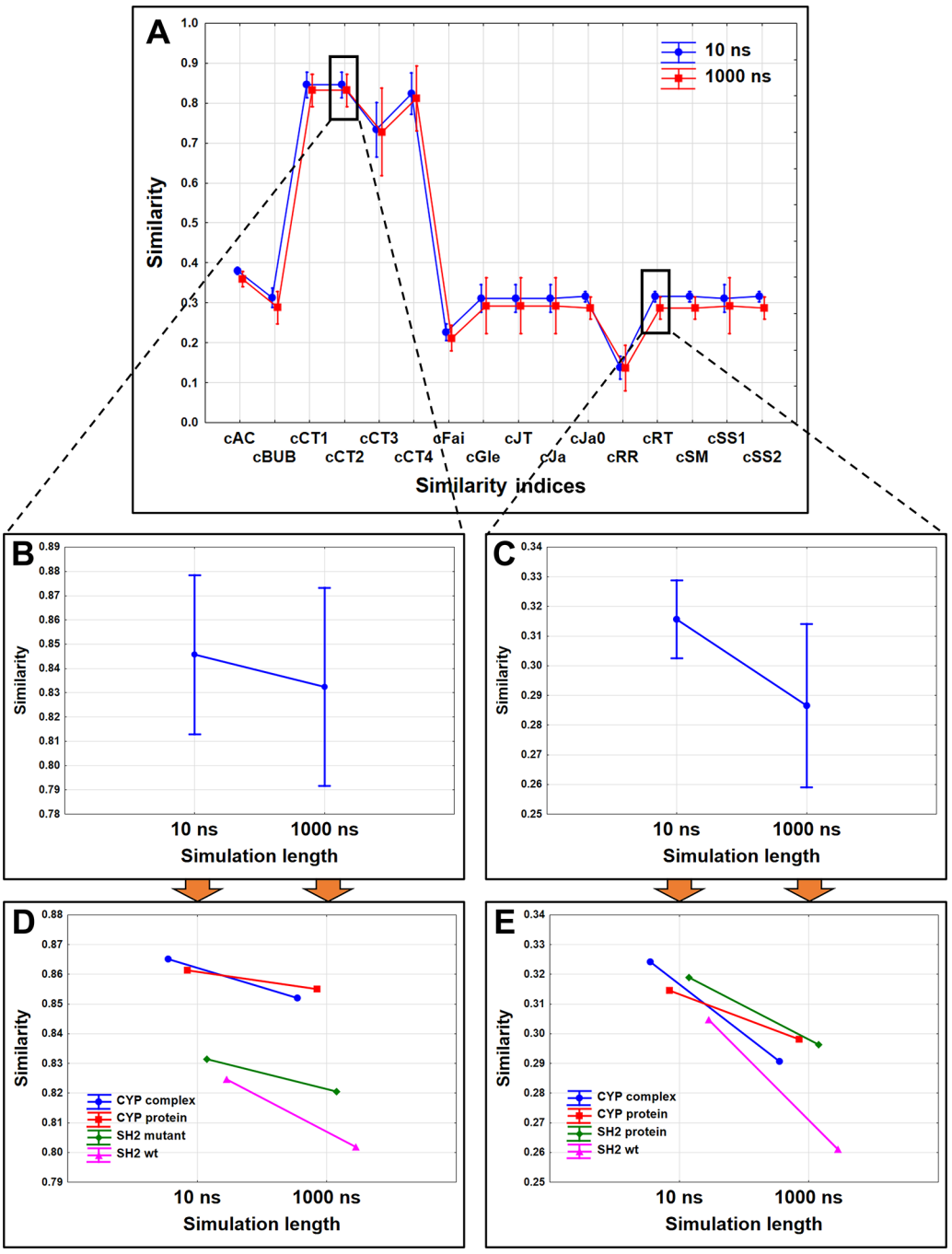


Figure 5. Extended continuous similarity values for the case studies merged: (A) based on the different similarity indices. cCT2 and cRT are emphasized in B and C, respectively, where the MD simulation length is compared. In the bottom row (D, E), the similarity of the frames for each case study is plotted separately in both simulation length cases. Similarities, plotted in the figure, are average similarity values with the 95% confidence intervals (except for the bottom graphs, where single similarity values are compared).

domain reveals three clusters, while the mutant SH2 domain assumes several recurring conformational states.

The coordinates extracted from the MD frames were used to calculate the 16 different, extended continuous similarity indices for each set of MD simulations. The similarity indices were merged into one data set, and factorial ANOVA was used for their comparison in combination with the simulation length as another factor (Figure 5). Due to the separation of the indices based on the range of the similarity values (y axis), we have presented the difference between the MD simulation lengths with the use of the extended continuous Consonni-Todeschini 2 and Rogers-Tanimoto (always in nonweighted versions, cCT2 and cRT, respectively) separately. As we observe the two mentioned indices more closely, the difference becomes more apparent: in both cases, the similarity between the frames is lower for the longer MD simulations, which is consistent with a more thorough sampling of the conformational space. Viewing the similarity of the case studies separately, we can see that the decrease of similarity is more pronounced for the SH2 domain simulations, especially for the wild-type protein. The result is in concordance with the findings in the t-SNE plots, where the

https://doi.org/10.1021/acs.jcim.2c00433
J. Chem. Inf. Model. 2022, 62, 3415−3425

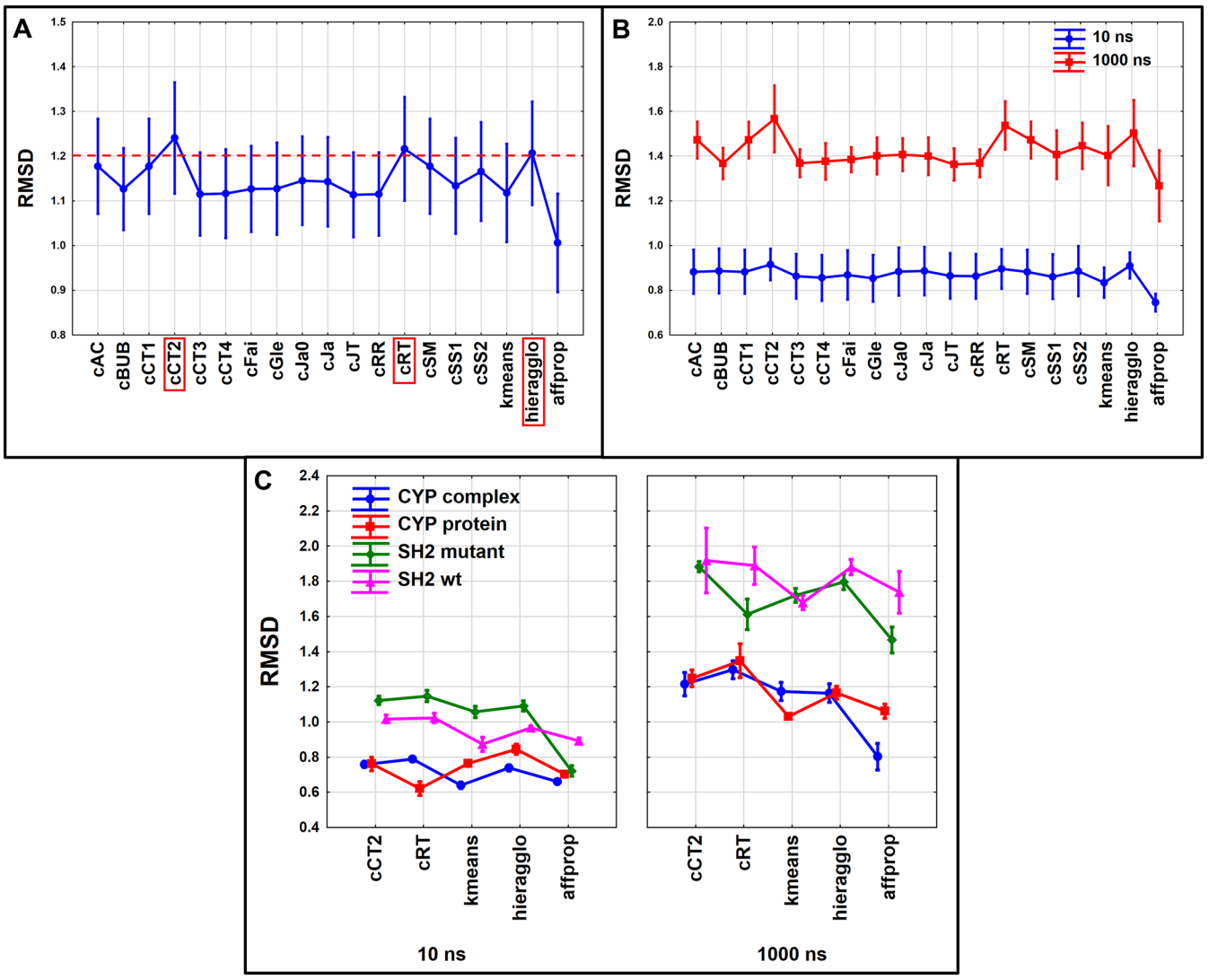


Figure 6. Comparison of the diversity selection methods with ANOVA. Average RMSD values of the selected frames are plotted on the *y* axis in every case. A) The red dotted line highlights the methods with the best performance (their names are highlighted on the *x* axis). B) The same results are presented with the simulation length being an additional factor in ANOVA. C) The case studies were added as a third factor for the analysis. (Only the two best indices are shown for clarity.) Average RMSD values with 95% confidence intervals are plotted in every figure.

smaller SH2 domains were shown to be more flexible than the larger CYP 2C9 protein.

3.2. Diversity Selection of the MD Frames. In this section, we show the application of the extended continuous similarity indices for diversity selection, namely, to select 5 to 10 diverse frames out of each MD trajectory. This is a common practice in modeling studies, producing diverse protein conformations, *e.g.*, for ensemble docking. We have compared the performance of our diversity picker to three clustering algorithms as benchmarks: (i) the affinity propagation algorithm implemented in the Schrödinger suite (from here on, affprop),[47] (ii) the hierarchical agglomeration approach (hieragglo), and (iii) the *k*-means method (kmeans) implemented in the AMBER tools simulation package. All 16 extended continuous indices were applied for the diversity selection with the ECS-MeDiv method. In each case, we had to select the most diverse 5−10 out of the 1000 frames of trajectory. Finally, root-mean-square distance (RMSD) values were calculated between each pair of the selected frames, and the RMSD values were compared with ANOVA. The merged data set for the ANOVA comparison contained the RMSD values of each selection case (5−10 diverse frame selection), with each similarity indices (16 + 3 benchmarks) and for all the four case studies. We used RMSD values as a performance parameter because it is the most commonly used indicator in MD and 3D molecular modeling in general. Here, larger RMSDs are better, since ideally our aim is to select structurally diverse frames to better represent the complete conformational space of the protein. In terms of RMSD, two extended continuous indices, namely cCT2 and cRT, were in a comparable range with the hieragglo benchmark algorithm, although the 95% confidence intervals were very wide in each case (Figure 6A). The reason for the wide confidence intervals is clear: the range of the RMSD values is different for the 10 and 1000 ns simulation lengths (Figure 6B), with the differences between the indices being larger at the longer simulation length, as we would expect. Therefore, we have included the simulation length as a factor in the next step, along with the similarity indices and the case studies. (We have selected the most promising cCT2- and cRT-based ECS-MeDiv algorithms in Figure 6C for clarity.) All the applied factors have

https://doi.org/10.1021/acs.jcim.2c00433
J. Chem. Inf. Model. 2022, 62, 3415−3425

carried statistically significant differences (α = 0.05), and in seven out of eight cases, cCT2 or cRT could beat all three benchmark methods. Variances are established based on the data points corresponding to the selection of 5−10 frames. Interestingly, the hieragglo algorithm outperformed the other two benchmark algorithms in each case.

As an additional comparison between the extended similarity indices and the clustering algorithms, we have rank-transformed the RMSD results and counted the "wins" (highest RMSD) for each method. The cCT2-based ECS-MeDiv method dominated this competition with 12 wins, followed by the cRT-based ECS-MeDiv and the hieragglo methods with 8−8 wins each. On the other hand, cRT has the lowest sum of rankings, which means that usually the cRT-based ECS-MeDiv selection could provide higher RMSD values for the selected frames than most of the other indices (see Supporting Information Table S1).

We also compared the calculation time of our two best-performing methods and the benchmark methods on the 100 μs SARS-CoV-2 main protease trajectory as a case study for a state-of-the-art long MD simulation. The affinity propagation algorithm was omitted as it showed poor performance in the previous studies. Due to highly increased calculation times, we have compared the results only for the selection of 10 frames out of 100,000. Table 1 shows the average RMSD values between the selected 10 frames for the cRT- and cCT2-based ECS-MeDiv selections along with two benchmark methods.

Table 1. RMSD Values and Runtimes for Frame Selection from the 3CLPro Long MD Case Study[b]

method	RMSD	std	time (h)[a]
kmeans	1.429	0.314	35.4
hieragglo	1.395	0.270	228.0
cCT2	1.605	0.495	**3.2**
cRT	**1.701**	0.514	**3.2**

[a]Performed on one thread of an AMD EPYC 7451 24-Core Processor. [b]std = standard deviation, see Materials and Methods.

Clearly, our extended continuous similarity-based diversity selection algorithms performed better, not only providing the best average RMSD values but also doing so in a drastically more efficient way. The calculations were 1 and 2 orders of magnitude faster than the benchmark kmeans and hieragglo algorithms, thanks to the linear (rather than quadratic) scaling of our method. Notice that while the hieragglo method came close to our algorithm in terms of RMSD values, it was the most time-demanding method.

4. CONCLUSIONS

We have presented the application of the extended continuous similarity indices as a novel tool to describe the conformational diversity in a set of structures, such as the frames of a molecular dynamics simulation. In addition, a new frame selection algorithm called ECS-MeDiv was developed, and its applicability and performance were demonstrated in diverse frame selection from molecular dynamics simulations and benchmarked against state-of-the-art solutions. The extended continuous similarity indices showed outstanding results in terms of RMSDs; two out of the 16, namely the cCT2 and cRT, indices could provide more diverse frame sets than the benchmark methods. Additionally, comparing the computational requirements for the frame selection from extremely long 100 μs SARS-CoV-2 main protease trajectory with 100,000 frames showed that the ECS-MeDiv, coupled with the cCT2 and cRT indices, drastically outperforms the hierarchical agglomerative method. In summary, the developed cCT2- and cRT-based ECS-MeDiv selection algorithms are suitable for the diverse selection of molecular structures extracted from MD trajectories. Moreover, the application of the extended continuous similarity indices is entirely general, and they can be easily adapted to describe the similarity of any data set with continuous values. The algorithm, along with basic usage examples, is available open-source at https://github.com/ramirandaq/MultipleComparisons.

5. DATA AND SOFTWARE AVAILABILITY

The Maestro molecular modeling program package and Desmond are commercial software with paid licenses. The AmberTools suite is free of charge, and its components are mostly released under the GNU General Public License (GPL). The molecular visualization software VMD is available to noncommercial users under a distribution-specific license. KNIME, the Konstanz Information Miner, is a free and open-source data analytics, reporting, and integration platform. Statistica is a proprietary advanced analytics software package. The source code of GNUplot, a portable command-line driven graphing utility, is copyrighted but freely distributed. Sample calculations and our scripts for the calculation of extended continuous similarities are available open-source at https://github.com/ramirandaq/MultipleComparisons.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jcim.2c00433.

Table of rank sums for two best extended continuous similarity indices, example calculation of nonweighted extended continuous Rogers-Tanimoto index, and table with formulas and notations for new, extended continuous similarity indices (PDF)

■ AUTHOR INFORMATION

Corresponding Authors

Károly Héberger − *Plasma Chemistry Research Group, Research Centre for Natural Sciences, 1117 Budapest, Hungary;* orcid.org/0000-0003-0965-939X; Email: heberger.karoly@ttk.hu

Ramón Alain Miranda-Quintana − *Department of Chemistry and Quantum Theory Project, University of Florida, Gainesville, Florida 32611, United States;* orcid.org/0000-0003-2121-4449; Email: quintana@chem.ufl.edu

Authors

Anita Rácz − *Plasma Chemistry Research Group, Research Centre for Natural Sciences, 1117 Budapest, Hungary;* orcid.org/0000-0001-8271-9841

Levente M. Mihalovits − *Medicinal Chemistry Research Group, Research Centre for Natural Sciences, 1117 Budapest, Hungary;* orcid.org/0000-0003-1022-3294

Dávid Bajusz − *Medicinal Chemistry Research Group, Research Centre for Natural Sciences, 1117 Budapest, Hungary;* orcid.org/0000-0003-4277-9481

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jcim.2c00433

Author Contributions

‖A.R. and L.M. contributed equally.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This work was supported under project No. OTKA K 134260, K 135150, and PD 134416 by the Ministry of Innovation and Technology of Hungary, from the National Research, Development and Innovation Fund, financed under the K and PD type funding scheme, respectively. The work of D.B. was supported by the János Bolyai Research Scholarship of the Hungarian Academy of Sciences and the ÚNKP-21-5 New National Excellence Program of the Ministry for Innovation and Technology. R.A.M.-Q. thanks the University of Florida for their support in the form of a startup grant.

■ REFERENCES

(1) Jász, Á.; Rák, Á.; Ladjánszki, I.; Cserey, G. Classical Molecular Dynamics on Graphics Processing Unit Architectures. *WIREs Comput. Mol. Sci.* **2020**, *10* (2), e1444.

(2) Götz, A. W.; Williamson, M. J.; Xu, D.; Poole, D.; Le Grand, S.; Walker, R. C. Routine Microsecond Molecular Dynamics Simulations with AMBER on GPUs. 1. Generalized Born. *J. Chem. Theory Comput.* **2012**, *8* (5), 1542−1555.

(3) De Vivo, M.; Masetti, M.; Bottegoni, G.; Cavalli, A. Role of Molecular Dynamics and Related Methods in Drug Discovery. *J. Med. Chem.* **2016**, *59* (9), 4035−4061.

(4) Lau, D.; Jian, W.; Yu, Z.; Hui, D. Nano-Engineering of Construction Materials Using Molecular Dynamics Simulations: Prospects and Challenges. *Compos. Part B Eng.* **2018**, *143*, 282−291.

(5) Bottaro, S.; Lindorff-Larsen, K. Biophysical Experiments and Biomolecular Simulations: A Perfect Match? *Science (80-.).* **2018**, *361* (6400), 355−360.

(6) Hollingsworth, S. A.; Dror, R. O. Molecular Dynamics Simulation for All. *Neuron* **2018**, *99* (6), 1129−1143.

(7) Orellana, L. Large-Scale Conformational Changes and Protein Function: Breaking the in Silico Barrier. *Front. Mol. Biosci.* **2019**, *6*, 117.

(8) Geng, H.; Chen, F.; Ye, J.; Jiang, F. Applications of Molecular Dynamics Simulation in Structure Prediction of Peptides and Proteins. *Comput. Struct. Biotechnol. J.* **2019**, *17*, 1162−1170.

(9) Mihalovits, L. M.; Ferenczy, G. G.; Keserű, G. M. Mechanistic and Thermodynamic Characterization of Oxathiazolones as Potent and Selective Covalent Immunoproteasome Inhibitors. *Comput. Struct. Biotechnol. J.* **2021**, *19*, 4486−4496.

(10) Mihalovits, L. M.; Ferenczy, G. G.; Keserű, G. M. Affinity and Selectivity Assessment of Covalent Inhibitors by Free Energy Calculations. *J. Chem. Inf. Model.* **2020**, *60* (12), 6579−6594.

(11) Ugur, I.; Schroft, M.; Marion, A.; Glaser, M.; Antes, I. Predicting the Bioactive Conformations of Macrocycles: A Molecular Dynamics-Based Docking Procedure with DynaDock. *J. Mol. Model.* **2019**, *25* (7), 197.

(12) Acharya, A.; Agarwal, R.; Baker, M. B.; Baudry, J.; Bhowmik, D.; Boehm, S.; Byler, K. G.; Chen, S. Y.; Coates, L.; Cooper, C. J.; Demerdash, O.; Daidone, I.; Eblen, J. D.; Ellingson, S.; Forli, S.; Glaser, J.; Gumbart, J. C.; Gunnels, J.; Hernandez, O.; Irle, S.; Kneller, D. W.; Kovalevsky, A.; Larkin, J.; Lawrence, T. J.; LeGrand, S.; Liu, S.-H.; Mitchell, J. C.; Park, G.; Parks, J. M.; Pavlova, A.; Petridis, L.; Poole, D.; Pouchard, L.; Ramanathan, A.; Rogers, D. M.; Santos-Martins, D.; Scheinberg, A.; Sedova, A.; Shen, Y.; Smith, J. C.; Smith, M. D.; Soto, C.; Tsaris, A.; Thavappiragasam, M.; Tillack, A. F.; Vermaas, J. V.; Vuong, V. Q.; Yin, J.; Yoo, S.; Zahran, M.; Zanetti-Polzi, L. Supercomputer-Based Ensemble Docking Drug Discovery Pipeline with Application to Covid-19. *J. Chem. Inf. Model.* **2020**, *60* (12), 5832−5852.

(13) Totrov, M.; Abagyan, R. Flexible Ligand Docking to Multiple Receptor Conformations: A Practical Alternative. *Curr. Opin. Struct. Biol.* **2008**, *18* (2), 178−184.

(14) Bottegoni, G.; Rocchia, W.; Rueda, M.; Abagyan, R.; Cavalli, A. Systematic Exploitation of Multiple Receptor Conformations for Virtual Ligand Screening. *PLoS One* **2011**, *6* (5), e18845.

(15) Cardoso, W. B.; Mendanha, S. A. Molecular Dynamics Simulation of Docking Structures of SARS-CoV-2 Main Protease and HIV Protease Inhibitors. *J. Mol. Struct.* **2021**, *1225*, 129143.

(16) Roe, D. R.; Cheatham, T. E. PTRAJ and CPPTRAJ: Software for Processing and Analysis of Molecular Dynamics Trajectory Data. *J. Chem. Theory Comput.* **2013**, *9* (7), 3084−3095.

(17) Bowers, K. J.; Chow, E.; Xu, H.; Dror, R. O.; P. Michael, E.; Gregersen, B. A.; Klepeis, J. L.; Kolossvary, I.; Moraes, M. A.; Sacerdoti, F. D.; Salmon, J. K.; Shan, Y.; Shaw, D. E. Scalable Algorithms for Molecular Dynamics Simulations on Commodity Clusters. In *SC '06: Proceedings of the 2006 ACM/IEEE Conference on Supercomputing*; Association for Computing Machinery: New York, NY, USA, 2006; DOI: 10.1109/SC.2006.54.

(18) Miranda-Quintana, R. A.; Bajusz, D.; Rácz, A.; Héberger, K. Extended Similarity Indices: The Benefits of Comparing More than Two Objects Simultaneously. Part 1: Theory and Characteristics†. *J. Cheminform.* **2021**, *13* (1), 32.

(19) Bajusz, D.; Miranda-Quintana, R. A.; Rácz, A.; Héberger, K. Extended Many-Item Similarity Indices for Sets of Nucleotide and Protein Sequences. *Comput. Struct. Biotechnol. J.* **2021**, *19*, 3628−3639.

(20) Rácz, A.; Dunn, T. B.; Bajusz, D.; Kim, T. D.; Miranda-Quintana, R. A.; Héberger, K. Extended Continuous Similarity Indices: Theory and Application for QSAR Descriptor Selection. *J. Comput. Aided. Mol. Des.* **2022**, *36*, 157.

(21) Miranda-Quintana, R. A.; Rácz, A.; Bajusz, D.; Héberger, K. Extended Similarity Indices: The Benefits of Comparing More than Two Objects Simultaneously. Part 2: Speed, Consistency, Diversity Selection. *J. Cheminform.* **2021**, *13* (1), 33.

(22) Flores-Padilla, E. A.; Juárez-Mercado, K. E.; Naveja, J. J.; Kim, T. D.; Alain Miranda-Quintana, R.; Medina-Franco, J. L. Chemoinformatic Characterization of Synthetic Screening Libraries Focused on Epigenetic Targets. *Mol. Inform.* **2022**, *41*, e2100285.

(23) Dunn, T. B.; Seabra, G. M.; Kim, T. D.; Juárez-Mercado, K. E.; Li, C.; Medina-Franco, J. L.; Miranda-Quintana, R. A. Diversity and Chemical Library Networks of Large Data Sets. *J. Chem. Inf. Model.* **2022**, *62*, 2186.

(24) Chang, L.; Perez, A.; Miranda-Quintana, R. A. Improving the Analysis of Biological Ensembles through Extended Similarity Measures. *Phys. Chem. Chem. Phys.* **2021**, *24* (1), 444−451.

(25) Miranda-Quintana, R.-A.; Cruz-Rodes, R.; Codorniu-Hernandez, E.; Batista-Leyva, A. J. Formal Theory of the Comparative Relations: Its Application to the Study of Quantum Similarity and Dissimilarity Measures and Indices. *J. Math. Chem.* **2010**, *47* (4), 1344−1365.

(26) Miranda-Quintana, R. A.; Kim, T. D.; Heidar-Zadeh, F.; Ayers, P. W. On the Impossibility of Unambiguously Selecting the Best Model for Fitting Data. *J. Math. Chem.* **2019**, *57* (7), 1755−1769.

(27) Miranda-Quintana, R. A.; Bajusz, D.; Rácz, A.; Héberger, K. Differential Consistency Analysis: Which Similarity Measures Can Be Applied in Drug Discovery? *Mol. Inform.* **2021**, *40* (7), 2060017.

(28) Liu, B. A.; Jablonowski, K.; Raina, M.; Arcé, M.; Pawson, T.; Nash, P. D. The Human and Mouse Complement of SH2 Domain Proteins—Establishing the Boundaries of Phosphotyrosine Signaling. *Mol. Cell* **2006**, *22* (6), 851−868.

(29) de Araujo, E. D.; Orlova, A.; Neubauer, H. A.; Bajusz, D.; Seo, H.-S.; Dhe-Paganon, S.; Keserű, G. M.; Moriggl, R.; Gunning, P. T. Structural Implications of STAT3 and STAT5 SH2 Domain Mutations. *Cancers (Basel).* **2019**, *11* (11), 1757.

(30) Wingelhofer, B.; Maurer, B.; Heyes, E. C.; Cumaraswamy, A. A.; Berger-Becvar, A.; de Araujo, E. D.; Orlova, A.; Freund, P.; Ruge, F.; Park, J.; Tin, G.; Ahmar, S.; Lardeau, C.-H.; Sadovnik, I.; Bajusz, D.; Keserű, G. M.; Grebien, F.; Kubicek, S.; Valent, P.; Gunning, P. T.;

https://doi.org/10.1021/acs.jcim.2c00433
J. Chem. Inf. Model. 2022, 62, 3415−3425

Moriggl, R. Pharmacologic Inhibition of STAT5 in Acute Myeloid Leukemia. *Leukemia* **2018**, *32* (5), 1135−1146.
(31) de Araujo, E. D.; Erdogan, F.; Neubauer, H. A.; Meneksedag-Erol, D.; Manaswiyoungkul, P.; Eram, M. S.; Seo, H.-S.; Qadree, A. K.; Israelian, J.; Orlova, A.; Suske, T.; Pham, H. T. T.; Boersma, A.; Tangermann, S.; Kenner, L.; Rülicke, T.; Dong, A.; Ravichandran, M.; Brown, P. J.; Audette, G. F.; Rauscher, S.; Dhe-Paganon, S.; Moriggl, R.; Gunning, P. T. Structural and Functional Consequences of the STAT5BN642H Driver Mutation. *Nat. Commun.* **2019**, *10* (1), 2517.
(32) Werck-Reichhart, D.; Feyereisen, R. Cytochromes P450: A Success Story. *Genome Biol.* **2000**, *1* (6), reviews3003.1.
(33) Zanger, U. M.; Schwab, M. Cytochrome P450 Enzymes in Drug Metabolism: Regulation of Gene Expression, Enzyme Activities, and Impact of Genetic Variation. *Pharmacol. Ther.* **2013**, *138* (1), 103−141.
(34) Wienkers, L. C.; Heath, T. G. Predicting in Vivo Drug Interactions from in Vitro Drug Discovery Data. *Nat. Rev. Drug Discovery* **2005**, *4* (10), 825−833.
(35) Rácz, A.; Bajusz, D.; Miranda-Quintana, R. A.; Héberger, K. Machine Learning Models for Classification Tasks Related to Drug Safety. *Mol. Divers.* **2021**, *25* (3), 1409−1424.
(36) Rácz, A.; Keserű, G. M. Large-Scale Evaluation of Cytochrome P450 2C9Mediated Drug Interaction Potential with Machine Learning-Based Consensus Modeling. *J. Comput. Aided. Mol. Des.* **2020**, *34* (8), 831−839.
(37) Swain, N. A.; Batchelor, D.; Beaudoin, S.; Bechle, B. M.; Bradley, P. A.; Brown, A. D.; Brown, B.; Butcher, K. J.; Butt, R. P.; Chapman, M. L.; Denton, S.; Ellis, D.; Galan, S. R. G.; Gaulier, S. M.; Greener, B. S.; De Groot, M. J.; Glossop, M. S.; Gurrell, I. K.; Hannam, J.; Johnson, M. S.; Lin, Z.; Markworth, C. J.; Marron, B. E.; Millan, D. S.; Nakagawa, S.; Pike, A.; Printzenhoff, D.; Rawson, D. J.; Ransley, S. J.; Reister, S. M.; Sasaki, K.; Storer, R. I.; Stupple, P. A.; West, C. W. Discovery of Clinical Candidate 4-[2-(5-Amino-1H-Pyrazol-4-Yl)-4-Chlorophenoxy]-5-Chloro-2-Fluoro-N-1,3-Thiazol-4-Ylbenzenesulfonamide (PF-05089771): Design and Optimization of Diaryl Ether Aryl Sulfonamides as Selective Inhibitors of NaV1.7. *J. Med. Chem.* **2017**, *60* (16), 7029−7042.
(38) Shaw, D. E.; Grossman, J. P.; Bank, J. A.; Batson, B.; Butts, J. A.; Chao, J. C.; Deneroff, M. M.; Dror, R. O.; Even, A.; Fenton, C. H.; Forte, A.; Gagliardo, J.; Gill, G.; Greskamp, B.; Ho, C. R.; Ierardi, D. J.; Iserovich, L.; Kuskin, J. S.; Larson, R. H.; Layman, T.; Lee, L.-S.; Lerer, A. K.; Li, C.; Killebrew, D.; Mackenzie, K. M.; Mok, S. Y.-H.; Moraes, M. A.; Mueller, R.; Nociolo, L. J.; Peticolas, J. L.; Quan, T.; Ramot, D.; Salmon, J. K.; Scarpazza, D. P.; Schafer, U. Ben; Siddique, N.; Snyder, C. W.; Spengler, J.; Tang, P. T. P.; Theobald, M.; Toma, H.; Towles, B.; Vitale, B.; Wang, S. C.; Young, C. Anton 2: Raising the Bar for Performance and Programmability in a Special-Purpose Molecular Dynamics Supercomputer. In *SC14: International Conference for High Performance Computing, Networking, Storage and Analysis*; IEEE: 2014; pp 41−53, DOI: 10.1109/SC.2014.9.
(39) Liu, C.; Zhou, Q.; Li, Y.; Garner, L. V.; Watkins, S. P.; Carter, L. J.; Smoot, J.; Gregg, A. C.; Daniels, A. D.; Jervey, S.; Albaiu, D. Research and Development on Therapeutic Agents and Vaccines for COVID-19 and Related Human Coronavirus Diseases. *ACS Cent. Sci.* **2020**, *6* (3), 315−331.
(40) Douangamath, A.; Fearon, D.; Gehrtz, P.; Krojer, T.; Lukacik, P.; Owen, C. D.; Resnick, E.; Strain-Damerell, C.; Aimon, A.; Ábrányi-Balogh, P.; Brandão-Neto, J.; Carbery, A.; Davison, G.; Dias, A.; Downes, T. D.; Dunnett, L.; Fairhead, M.; Firth, J. D.; Jones, S. P.; Keeley, A.; Keserü, G. M.; Klein, H. F.; Martin, M. P.; Noble, M. E. M.; O'Brien, P.; Powell, A.; Reddi, R. N.; Skyner, R.; Snee, M.; Waring, M. J.; Wild, C.; London, N.; von Delft, F.; Walsh, M. A. Crystallographic and Electrophilic Fragment Screening of the SARS-CoV-2 Main Protease. *Nat. Commun.* **2020**, *11* (1), 5047.
(41) Owen, D. R.; Allerton, C. M. N.; Anderson, A. S.; Aschenbrenner, L.; Avery, M.; Berritt, S.; Boras, B.; Cardin, R. D.; Carlo, A.; Coffman, K. J.; Dantonio, A.; Di, L.; Eng, H.; Ferre, R.; Gajiwala, K. S.; Gibson, S. A.; Greasley, S. E.; Hurst, B. L.; Kadar, E. P.; Kalgutkar, A. S.; Lee, J. C.; Lee, J.; Liu, W.; Mason, S. W.; Noell, S.; Novak, J. J.; Obach, R. S.; Ogilvie, K.; Patel, N. C.; Pettersson, M.; Rai, D. K.; Reese, M. R.; Sammons, M. F.; Sathish, J. G.; Singh, R. S. P.; Steppan, C. M.; Stewart, A. E.; Tuttle, J. B.; Updyke, L.; Verhoest, P. R.; Wei, L.; Yang, Q.; Zhu, Y. An Oral SARS-CoV-2 M pro Inhibitor Clinical Candidate for the Treatment of COVID-19. *Science (80-.).* **2021**, *374* (6575), 1586−1593.
(42) Madhavi Sastry, G.; Adzhigirey, M.; Day, T.; Annabhimoju, R.; Sherman, W. Protein and Ligand Preparation: Parameters, Protocols, and Influence on Virtual Screening Enrichments. *J. Comput. Aided. Mol. Des.* **2013**, *27* (3), 221−234.
(43) Harder, E.; Damm, W.; Maple, J.; Wu, C.; Reboul, M.; Xiang, J. Y.; Wang, L.; Lupyan, D.; Dahlgren, M. K.; Knight, J. L.; Kaus, J. W.; Cerutti, D. S.; Krilov, G.; Jorgensen, W. L.; Abel, R.; Friesner, R. A. OPLS3: A Force Field Providing Broad Coverage of Drug-like Small Molecules and Proteins. *J. Chem. Theory Comput.* **2016**, *12* (1), 281−296.
(44) Nosé, S. A Molecular Dynamics Method for Simulations in the Canonical Ensemble. *Mol. Phys.* **1984**, *52* (2), 255−268.
(45) Hoover, W. G. Canonical Dynamics: Equilibrium Phase-Space Distributions. *Phys. Rev. A* **1985**, *31* (3), 1695−1697.
(46) Humphrey, W.; Dalke, A.; Schulten, K. VMD: Visual Molecular Dynamics. *J. Mol. Graph.* **1996**, *14* (1), 33−38.
(47) Frey, B. J.; Dueck, D. Clustering by passing Messages Between Data Points. *Science (80-.).* **2007**, *315* (5814), 972−976.
(48) Zepeda-Mendoza, M. L.; Resendis-Antonio, O. Hierarchical Agglomerative Clustering. In *Encyclopedia of Systems Biology*; Springer New York: New York, NY, 2013; pp 886−887, DOI: 10.1007/978-1-4419-9863-7_1371.
(49) Jin, X.; Mannor, S.; Han, J.; Zhang, X. K-Means Clustering. In *Encyclopedia of Machine Learning*; Springer US: Boston, MA, 2011; pp 563−564, DOI: 10.1007/978-0-387-30164-8_425.
(50) van der Maaten, L.; Hinton, G. Visualizing Data Using T-SNE. *J. Mach. Learn. Res.* **2008**, *9* (86), 2579−2605.
(51) Berthold, M. R.; Cebron, N.; Dill, F.; Gabriel, T. R.; Kötter, T.; Meinl, T.; Ohl, P.; Thiel, K.; Wiswedel, B. KNIME - the Konstanz Information Miner. *ACM SIGKDD Explor. Newsl.* **2009**, *11* (1), 26−31.
(52) GNUplot. http://www.gnuplot.info/ (accessed 2022-07-06).

https://doi.org/10.1021/acs.jcim.2c00433
J. Chem. Inf. Model. 2022, 62, 3415−3425