JCTC Journal of Chemical Theory and Computation

pubs.acs.org/JCTC Article

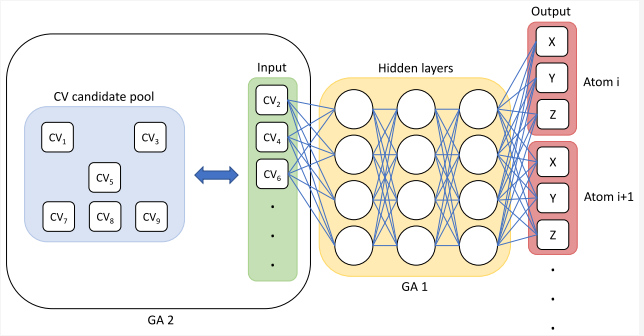
Discovering Collective Variables of Molecular Transitions via Genetic Algorithms and Neural Networks

Ferry Hooft, Alberto Pérez de Alba Ortíz, and Bernd Ensing*

Cite This: J. Chem. Theory Comput. 2021, 17, 2294−2306 Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: With the continual improvement of computing hardware and algorithms, simulations have become a powerful tool for understanding all sorts of (bio)molecular processes. To handle the large simulation data sets and to accelerate slow, activated transitions, a condensed set of descriptors, or collective variables (CVs), is needed to discern the relevant dynamics that describes the molecular process of interest. However, proposing an adequate set of CVs that can capture the intrinsic reaction coordinate of the molecular transition is often extremely difficult. Here, we present a framework to find an optimal set of CVs from a pool of candidates using a combination of artificial neural networks and genetic algorithms. The approach effectively replaces the encoder of an autoencoder network with genes to represent the latent space, i.e., the CVs. Given a selection of CVs as input, the network is trained to recover the atom coordinates underlying the CV values at points along the transition. The network performance is used as an estimator of the fitness of the input CVs. Two genetic algorithms optimize the CV selection and the neural network architecture. The successful retrieval of optimal CVs by this framework is illustrated at the hand of two case studies: the well-known conformational change in the alanine dipeptide molecule and the more intricate transition of a base pair in B-DNA from the classic Watson−Crick pairing to the alternative Hoogsteen pairing. Key advantages of our framework include the following: optimal interpretable CVs, avoiding costly calculation of committor or time-correlation functions, and automatic hyperparameter optimization. In addition, we show that applying a time-delay between the network input and output allows for enhanced selection of slow variables. Moreover, the network can also be used to generate molecular configurations of unexplored microstates, for example, for augmentation of the simulation data.

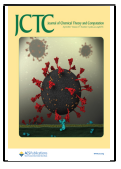


INTRODUCTION

Molecular dynamics (MD) simulations can provide unique insight into the structure, the dynamics, and the thermodynamic properties of molecular systems. However, many molecular processes of interest are activated and take place on time scales that are much longer than accessible by straightforward MD simulation. Examples include chemical reactions, phase transitions, protein folding, and self-assembly processes. To overcome this so-called "rare event" problem, a plethora of enhanced sampling methods has been developed throughout the years, such as umbrella sampling,[1] constrained MD,[2] steered MD,[3] metadynamics,[4] and transition path sampling (TPS),[5] to name but a few. Many of these methods apply a bias to the dynamics of the molecular system during the simulation along one or several well-chosen geometric descriptors of the transition, or "collective variables (CVs)", allowing the system to cross activation barriers and to sample the transitions of interest. Certain enhanced sampling techniques can map the underlying free energy landscape in the area spanned by the CVs, including its stable states and separating bottlenecks, from which transition rates and equilibrium constants are obtained. Other techniques aim to probe the reaction paths, which yields insight into the transition mechanism. Almost always the performance of the enhanced sampling technique depends critically on the choice of the CVs, and a suboptimal description of the transition is usually detrimental for obtaining statistically converged properties. In addition, CVs are also useful for deciphering and interpreting the high-dimensional information from the generated trajectories. Finding these CVs is thus a paramount objective in computational molecular science. Unfortunately, choosing CVs usually relies on human intuition, good chemical insight, and trial and error and is fraught with difficulty for all but the simplest transitions.[6,7]

The availability of large data sets from unbiased or enhanced simulations has led to development of various automated

Received: September 23, 2020
Published: March 4, 2021



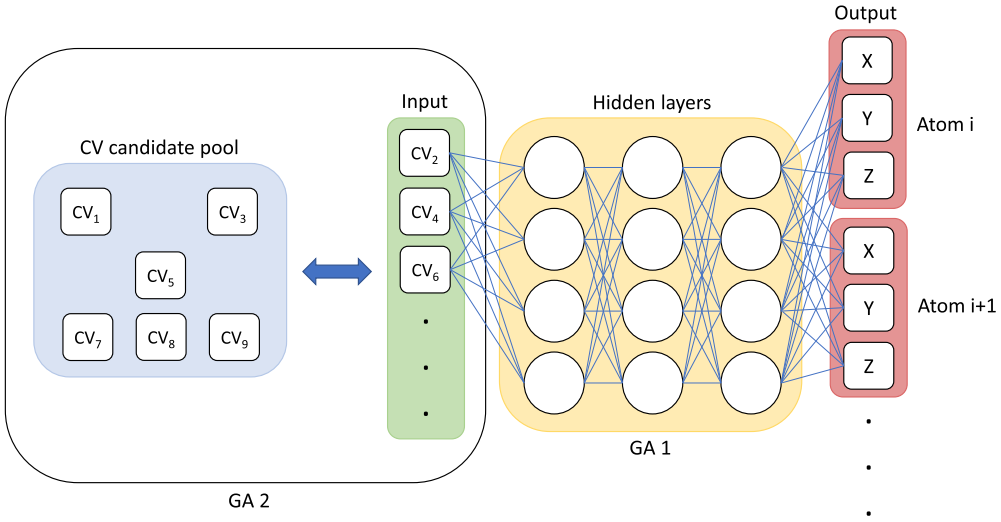


Figure 1. Schematic view of the framework. The inputs (green) of the network are a subset of the candidate CVs. GA 2 optimizes both the number of inputs and decides which of the candidate CVs out of the candidate pool (blue) are best for the lowest output error with respect to the entire training pool. GA 1 optimizes the hidden layer architecture. The output of the network (red) is in the XYZ-format and can consist of the current-time t coordinates or the lagged-time $t + \tau$ coordinates. An individual's genes consist of an input layer and the hidden layers, which are internally represented using a Python dictionary.[46] During the course of the two optimization steps, only the output remains constant; the number of atoms in the systems does not change.

methods for finding CVs using techniques from data science, information theory, and machine learning, such as singular value decomposition,[8] principle component analysis,[9−11] linear discriminant analysis,[12,13] maximum likelihood estimation,[14,15] and cross-entropy minimization.[16] In 2005, Ma and Dinner were the first to automate the selection of optimal CVs from a large set of predefined descriptors using a combination of two machine learning techniques: (1) a genetic algorithm (GA)[17] would evolve a pool of small sets of CVs, randomly chosen from the large CV set, by crossbreeding and random mutations and (2) an artificial neural network (NN)[18] would determine the quality, or "fitness", of each set of CVs as its ability to predict the progress along a reaction coordinate.[19] Here, a reference for the reaction progress was computed beforehand for a large data set of molecular configurations as the so-called committor probability of subsiding into the product state, by starting many MD trajectories with random velocities from each configuration and counting how many arrived in the reactant and product states. The NN, typically represented as a network of connected nodes, with an input layer, one or more hidden layers, and an output layer (see, for example, Figure 1), functions as a highly parametrized, and therefore very versatile, fit function, which was parametrized, or "trained", by supplying many CV values from MD configurations as input and comparing the NN output to the expected reaction coordinate (i.e., the committor) value.

Since then, several advanced techniques have been developed for CV discovery using NNs. For example, to avoid the calculation of the computationally costly committor probabilities to measure the quality of the CVs, Noé and co-workers developed a method based on lagged-time independent component analysis (TICA), which selects CVs that are optimal for generating the dominant eigenvalues of a transition rate matrix.[20] VAMPnets, developed in the same group, is an extension of this principle that uses NNs.[21] Instead of selecting CVs from a prior list of candidates, an autoencoder (i.e., an NN with a node structure shaped as an hourglass) can be employed that takes atom coordinates for both input and output and passes them through a bottleneck, or "latent space", in the network (i.e., the narrow middle layer of the "hourglass") that optimizes to a reduced manifold representing the slow modes.[22−24] The second "encoder" half of such networks can be used, for example, to generate unexplored configurations for further sampling, allowing for data generation and CV discovery in an iterative manner.[25−28] Also, the SGOOP method by Tiwary and Berne uses iterations of enhanced simulations to find CVs, but instead of an NN, it optimizes the time scale separation, or spectral gap, between the slow and fast eigenvalues of the transition probability matrix.[29] Finally, three rather different approaches, not using NNs, for finding a reduced representation of the slow process are on-the-fly generation of low-dimensional maps, such as SketchMap[30] and diffusion maps,[31,32] methods that map the free energy of a small set of CVs using an adiabatic separation and increase of the temperature of the CV subset, such as AFED,[33] TAMD,[34] and d-AFED,[35] and methods that on-the-fly optimize a nonlinear string[36,37] or path-CV[38−40] in the space of predefined CVs. Of the latter approach is also an NN-based implementation.[41]

Note that automated methods for finding CVs require as input data sets of trajectories that sample the molecular transition, which makes it a chicken and egg problem, because sampling the transition with an enhanced sampling technique typically requires good CVs. In practice, obtaining a satisfactory set of CVs and a converged sampling over a transition barrier is often an iterative process of trial-and-error until self-consistency. To obtain an initial data set of reactive trajectories, one could, for example, steer the molecular dynamics along a simple (set of) CV(s), or one could raise the temperature to overcome the free energy barrier. In this work, we will use data obtained with transition path sampling simulations, which is an enhanced sampling technique that only requires relatively simple order parameters that define the stable states. The resulting CVs can then be used, for example, to compute free energy landscapes.

https://dx.doi.org/10.1021/acs.jctc.0c00981
J. Chem. Theory Comput. 2021, 17, 2294−2306

In this work, we present a framework that combines several of the above ideas. Similar to the original work of Dinner and Ma, our framework is able to select the optimal CVs from a pool of candidate CVs. This guarantees intuitive human-interpretable descriptors of the molecular process. Examples of candidate CVs include the following: distances between atoms, angles, dihedral angles, more collective geometric coordinates, such as coordination numbers, contact maps, distances between the centers of masses of atom groups, and so forth. From this pool, CVs are selected and used as input to an NN. For the NN output, instead of using computationally demanding committor functions or time-correlation functions, we simply use the atomic coordinates. The optimal set of CVs is therefore the one that performs best in reconstructing the molecular configurations given as NN input the CV values obtained from those configurations. Configurations from MD simulation trajectories can thus be directly used as targets, or "labels", for supervised learning. Moreover, after training, the NN can be used to generate new configurations of still unexplored spatial regions, to launch additional MD simulations for further sampling. The individual CVs are given a fitness score based on the accuracy of the recreated coordinates. This score is then used by a GA to select the fittest combination of CVs. In doing so, the GA is effectively selecting CVs that best represent the transition in the trajectories.[42]

In the following, we first describe our machine learning framework to find optimal CVs from a large set of candidate CVs. The method is illustrated at the hand of two conformational transitions: (1) the well-known $C7_{ax}$-to-$C7_{eq}$ transition in the prototypical alanine dipeptide molecule and (2) the base-pair flipping of an adenine nucleic acid in a short segment of double-stranded B-DNA from the canonical Watson−Crick (WC) pairing to the less stable Hoogsteen (HG) configuration. We end with conclusions and an outlook.

METHODS

We will first describe our combined GA/NN framework and then introduce the two molecular systems that we use as case studies, the $C7_{ax}$-to-$C7_{eq}$ transition in alanine dipeptide and the WC-to-HG base rolling in DNA, together with the simulation details that were used to generate the data. The final subsection provides details on the generation of the input data sets, i.e., the lists of CVs, for both case studies.

The Framework for Finding Important Collective Variables. In our framework, schematically drawn in Figure 1, the selection of CVs from a pool of candidates is optimized using a genetic algorithm (here denoted "GA 2") by attributing a fitness to each CV that quantifies its ability to predict the coordinates of the atoms in the system. This fitness is computed using the NN (the green, yellow, and red parts in Figure 1), after training, and taken as the average error in the output coordinates predicted by the NN with respect to the actual coordinates used to generate the input CV values. The error is averaged over a large set of molecular configurations that sample the molecular transition under investigation, obtained by MD simulation. The lower the average error, the higher the fitness of the CVs in the current set.

The goal of the framework is to find the set of CVs, out of the list of candidate CVs, that can best recreate the coordinates of the system at any point along the transition. In other words, we try to find which CVs yield the lowest error when used as input to an NN that tries to reproduce the coordinates of trajectories that include a transition. One may note that the length of the NN output vector, i.e., the number of atom position coordinates (indicated in red in Figure 1), will almost always be larger than the NN input vector, i.e., the number of CVs (green in Figure 1). When not familiar with *undercomplete* autoencoder networks, in which the output dimension is always larger than the dimension of the feature space that is the input of the encoder half, this may seem intuitively a badly underdetermined problem. In practice however, this is not a problem because, in the first place, the output coordinates are not independent variables but are related via the interactions in the structure (e.g., bonds, bends, dihedrals, etc.). Second, even without providing input CVs (just providing zeroes as input, say), the final layer may simply learn an average structure from the training data (the frames are typically already aligned by translation and rotation), which is already a reasonable prediction of the coordinates. Therefore, the network only needs to learn how each input CV *changes* the coordinates along the molecular transition. Of course, there can be additional, independent motions in atoms (e.g., from solvent molecules or side group rotamers) that the network cannot predict with the given input CVs, but this only adds some irrelevant noise that is the same for all trial CV sets. Moreover, the user can omit redundant atom coordinates from the output vector, as we will illustrate in the second case study.

The optimization process is divided into two separate steps carried out by two genetic algorithms:

I GA 1: Optimizes the hidden layer architecture of the NN for optimal reproduction of the atomic coordinates, while keeping the inputs, typically including the entire CV pool, fixed.

II GA 2: Optimizes the number of CVs and the selection of CVs that are used as network inputs, while keeping the hidden layers of the NN fixed at the NN architecture determined in step 1 by GA 1.

This two-step approach is implemented to eliminate the bottleneck that can be caused by a suboptimal network architecture and relies on the idea that a network architecture optimized for all CVs may also accommodate training based on a subset of these features. Although different approaches to a similar feature selection exist, this approach allows for a large pool of candidates without *a priori* knowledge of the search space.[43−45] It has been divided into two steps to limit the search space and ease the interpretation of the results.

A second variant of the framework was created by including a so-called lag-time, as previously done, for example, by Wehmeyer et al.,[47] resulting in a method to distinguish and favor the slowest CVs. The difference between this "lagged-time" variant with the original "current-time" variant is entirely in the output of the network (indicated in red in Figure 1): the atomic coordinates at time $t + \tau$ are inferred from the CVs at time t instead of the atomic coordinates at time t. In other words, the NN of the lagged-time variant takes as input CV values from a configuration at time t and aims to predict as output what the atomic coordinates should be some lag-time τ later.

For training and testing of the NN, the data set is divided into three different sets: the training set, the validation set, and the testing set. The first is used to train the network weights, the second is used to monitor and stop the training (early stopping) to avoid overfitting, and the last is used to determine the final performance of the trained network. Training has

https://dx.doi.org/10.1021/acs.jctc.0c00981
J. Chem. Theory Comput. 2021, 17, 2294−2306

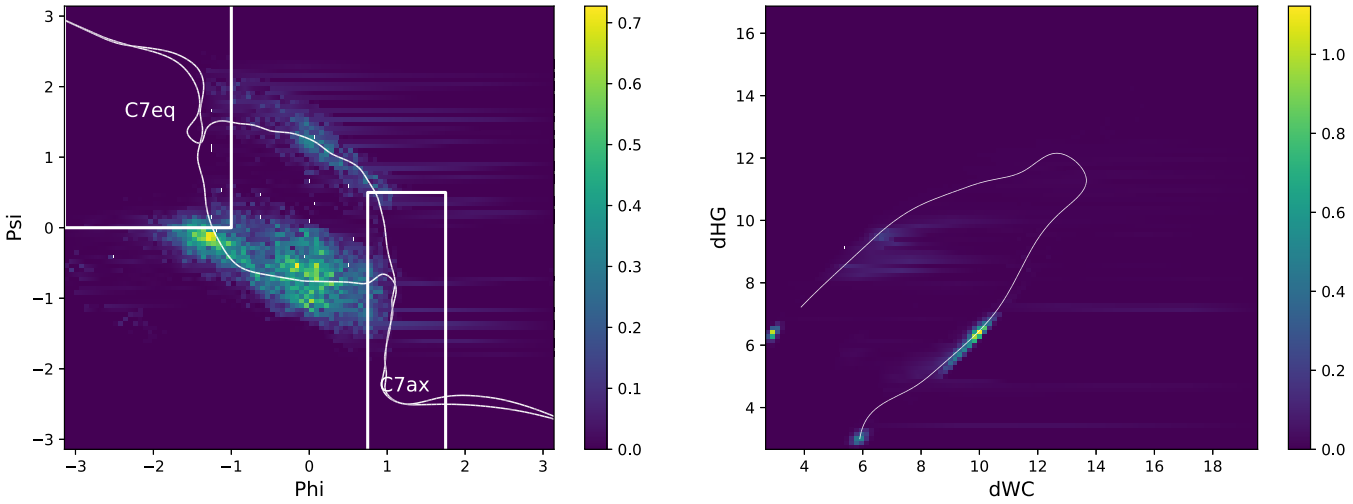


Figure 2. Left: Normalized path density histogram of the TPS simulation of alanine dipeptide. A lighter color (yellow) indicates a higher concentration of visits to this particular configuration of CVs during all combined transitions. The stable reactant and product states are indicated by white rectangles. Note that there are two mechanistic channels, also indicated by the thin white lines marking the mean transition paths. **Right:** Normalized path density histogram of the DNA transition with respect to the distances dWC and dHG, found in Table 4 and Figure 3B.

been performed using the Keras library;[48] details on the global settings of the NN optimization are provided in the Supporting Information.

At the start of the genetic algorithms, a "pool of individuals" is created based on parameter constraints that have been established by the user (see the Supporting Information for an example). In optimization step 1, each individual consists of the entire pool of CVs and a hidden layer architecture, corresponding to the green and yellow fields of Figure 1, respectively. In step 2, each individual consists of a subset of CVs and the optimized hidden layer architecture from step 1. In each step, the NNs of the individuals are trained and assigned a fitness score that is based on their mean absolute error (MAE) loss when applied to a test data set and ranked. The top 10% best performing networks are allowed to transfer their genetic information to the next generation by entering the parent pool. From the rejects, a further 2% is randomly selected to enter the genetic pool to increase diversity. From the genetic pool, two parents are randomly selected with equal weighting for every individual in the pool, in order to create two different children by crossover. After the creation of each child, a 10% chance exists for mutations to occur. In order to limit the number of CVs selected by the GA, a penalty p_{input} has been introduced that increases the final fitness of the network with a multiplication factor. Further details of the genetic algorithms are found in the Supporting Information.

The open-source Python-based implementation of this framework is freely available on our github website,[49] together with the data of the alanine dipeptide case study.

Molecular Systems and Simulation Setup. To illustrate the performance of our combined GA/NN framework for finding good CVs, we introduce in the following two biomolecular systems that undergo a conformational transition, which we will use as case studies. To obtain data sets of molecular configurations, we use classical MD trajectories obtained with the transition path sampling (TPS) technique[5] to enhance the sampling of the otherwise prohibitively rare transition. In brief, TPS is a stochastic (Monte Carlo-based) algorithm to generate new reactive trajectories from an initial one, by randomly selecting a configuration, making small random changes to all atomic velocities, and using this as a starting point to branch off a new MD trajectory. The new trajectory is rejected or accepted based on a Metropolis criterion, after which, if accepted, it is used as the starting point for the next trajectory, and so forth. The generated ensemble of transition trajectories (or "paths") is used for statistical analysis of the molecular transition or, in our case, to find optimal CVs that describe the transition.

Alanine Dipeptide. The $C7_{ax}$-to-$C7_{eq}$ transition of alanine dipeptide was sampled using the OpenPathSampling package[50,51] to perform flexible-length TPS with a one-way shooting move scheme and uniform selection. Figure 2 shows the resulting path density histogram, describing the average transition path between the two states. For the purpose of rapid data generation, the two states were defined larger than conventional by setting the boundaries of the torsion collective variables ϕ and ψ as follows:

$$C7_{eq}\begin{cases} -\pi \le \phi \ge -1 \\ 0 \le \psi \ge \pi \end{cases} \quad (1)$$

$$C7_{ax}\begin{cases} 0.75 \le \phi \ge 1.75 \\ -\pi \le \psi \ge 0.5 \end{cases} \quad (2)$$

The simulations were carried out using the OpenMM engine[52] with the amber99sb-ildn force field[53] and the velocity Verlet with velocity randomization (VVVR) integrator[54] at a temperature of 300 K with a collision rate of 1.0 ps^{-1} and a time step of 2.0 fs. The particle mesh Ewald (PME) method was employed with a cutoff of 1 nm.[55] No explicit or implicit solvation was used, and a frame was recorded every 10 time steps. From the generated paths, only the accepted paths were included in the final data set (see also Table 1).

B-DNA. TPS data of a Watson−Crick (WC)-to-Hoogsteen (HG) transition of the A16−T9 base pair in a segment of B-DNA was obtained from Vreede et al.[56] The transition of this base pair within the sequence 5-CGATTTTTTGGC-3 (complementary strand 3-GCTAAAAAACCG-5) has been studied in refs 57−59. In DNA, going from the WC-to-HG base pairing geometry, the flip of the purine (from anti in WC to syn in HG) is accomplished by a 180° rotation of the base

https://dx.doi.org/10.1021/acs.jctc.0c00981
J. Chem. Theory Comput. 2021, 17, 2294−2306

Table 1. Statistical Details of the Data Set Generation Using TPS Simulation for the Alanine Dipeptide Transition

parameter	value
number of accepted paths	4025
acceptance ratio	0.80
number of decorrelated paths	768
total number of snapshots	81458
total accepted snapshots	67577

Table 2. Statistical Details of the Data Set Generation Using TPS Simulation for the DNA Transition

parameter	value
accepted paths	1716
acceptance ratio	0.28
decorrelated paths	125
selected paths	300
selected snapshots	290937

along the bond connecting the base to the sugar, known as the glycosidic bond, leading to an alternative hydrogen bonding motif. This so-called "rolling" of the base can be accompanied by "flipping" outside of the confines of the double helix, as illustrated in Figure 3A. In this work, we focus on the transition with flipping, as it presents a larger conformational change.

For the TPS stable state definitions, Vreede and co-workers used the differences in hydrogen bond patterns between the WC and HG states. The system is considered to be in the WC state if both distances dHB, between atom O4 in residue DT9 and atom N6 in residue 4DA, and dWC (see Table 4 and Figure 3B for a description of these CVs) are below 0.35 nm, while the system is considered to be in the HG state if both distances dHB and dHG are below 0.35 nm. The cutoff of 0.35 nm is based on the maximum distance between a hydrogen bond donor and acceptor in a hydrogen bond. The flexible length TPS simulations started from a transition with the adenine in residue 4DA going outside the double helix into the solvent during the rotation. The B-DNA sequence is modeled with the BSC1 force field.[60] Further details regarding the simulation can be obtained from ref 56.

From all generated data, only a random selection of trajectories was used to limit the total amount of data. The final data set consisted of evenly spaced trajectories (based on their generation order) of all 10 runs. Statistical details are given in Table 2. Figure 2 shows the path density histogram of the transition with respect to the distances dHG and dWC (see Table 4 and Figure 3B).

Collective Variables. From the generated trajectories of both systems, a pool of CVs was extracted using the PLUMED 2.6.0[61,62] software. The CV candidates range from previously tested and accepted to questionably relevant for the transition of interest. For our purposes, a "good CV" refers to a descriptive slow coordinate of the transition in question that can be biased in accelerated sampling; in the case of alanine dipeptide, the good CVs refer to the known ϕ and ψ dihedral angles. Table 3 describes all candidate CVs of the alanine dipeptide system that were used to create the candidate pool. The omega and theta angles were selected to include more of the backbone dihedrals; the end-to-end distance as well as the NCaR-angle were selected to include CVs that relate to the global structure of the molecule; and finally, the dOO distance

Table 3. Candidate Collective Variables of the Alanine Dipeptide System and Their Descriptions Using the CHARMM Nomenclature

CV	type	description
phi	torsion	dihedral angle in the backbone formed by the axis NL-CA and the vectors CLP-NL and CA-CRP
psi	torsion	dihedral angle in the backbone formed by the axis CA-CRP and the vectors NL-CA and CRP-NR
omega	torsion	dihedral angle in the backbone formed by the axis CRP-NR and the vectors CA-CRP and NR-CR
theta	torsion	dihedral angle in the backbone formed by the axis CLP-NL and the vectors CL-CLP and NL-CA
NCaR_angle	angle	angle formed by the atoms NL, CA, and CB
end_Ca_end	angle	angle formed by the atoms CL, CA, and CR
dOO	distance	distance between atoms OL and OR

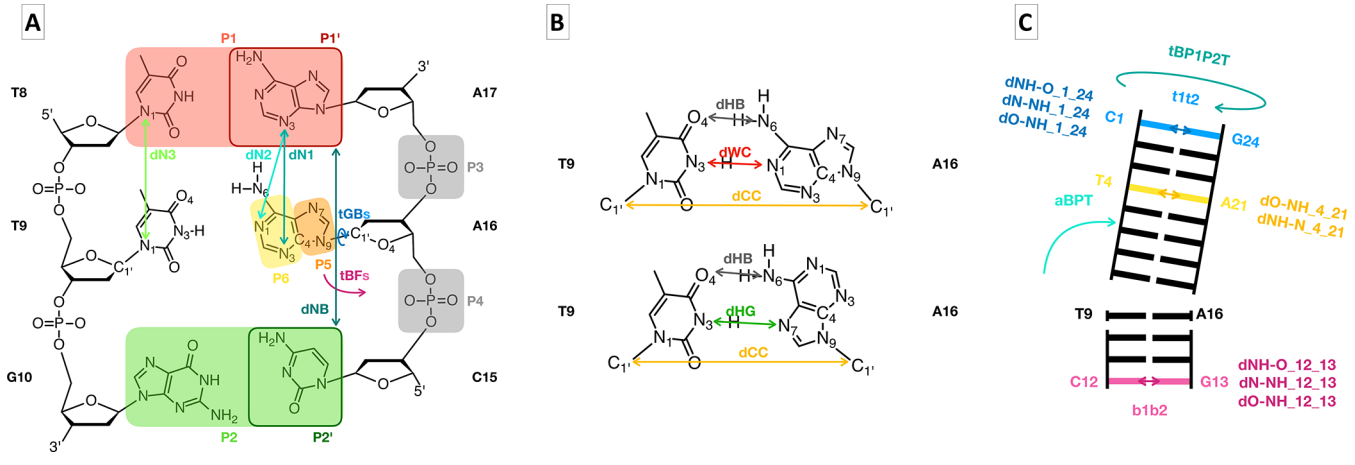


Figure 3. A: The A16−T9 base pair and its two neighboring base pairs above and below, with graphical representations of the centers of mass involved in the calculation of CVs dNB, tGB, and tBF. The distance CVs dN1, dN2, and dN3 and the dihedrals tGBs and tBFs are indicated as well. **B:** Top: Watson−Crick base pair with graphical representations of the CVs: dWC, dHB, and dCC; Bottom: Hoogsteen base pair with graphical representations of the CVs: dHG, dHB, and dCC. **C:** Schematic of the double-stranded DNA segment highlighting the base pairs involved in the distances dNH-O_12_13 and dNH-N_4_21 as well as the centers of mass: aBPT and tBP1P2T. Based on ref 59.

https://dx.doi.org/10.1021/acs.jctc.0c00981
J. Chem. Theory Comput. 2021, 17, 2294−2306

Table 4. Candidate Collective Variables of the DNA System and Their Descriptions[a]

CV	type	description
dWC	distance	distance of the H-bond between A16 (N1) and T9 (N3), characteristic for Watson−Crick
dHG	distance	distance of the H-bond between A16 (N7) and T9 (N3), characteristic for Hoogsteen
dHB	distance	distance of the H-bond between A16 (N6) and T9 (O4), present both in WC and in HG
dCC	distance	distance between A16 (C1) and T9 (C1)
dNB	distance	distance between the centers of mass P1 and P2
dN1	distance	distance between A16 (N3) and A17 (N3)
dN2	distance	distance between A16 (N1) and A17 (N3)
dN3	distance	distance between T8 (N1) and T9 (N1)
dNH-O_12_13	distance	distance between C12 (N4) and G13 (O6)
dN-NH_12_13	distance	distance between C12 (N3) and G13 (N1)
dO-NH_12_13	distance	distance between C12 (O2) and G13 (N2)
dNH-O_1_24	distance	distance between C1 (N4) and G24 (O6)
dN-NH_1_24	distance	distance between C1 (N3) and G24 (N1)
dO-NH_1_24	distance	distance between C1 (O2) and G24 (N2)
dO-NH_4_21	distance	distance between T4 (O4) and A21 (N6)
dNH-N_4_21	distance	distance between T4 (N3) and A21 (N1)
aBPT	angle	angle of the entire DNA segment with the A16−T9 base pair as the vertex, defined by the centers of mass b1b2, p_all, and t1t2
tGB	torsion	torsion around the glycosidic bond defined by the pseudodihedral angle formed by the axis A16 (C1−N9) and the vectors P2−P1 and P5−P6
tBF	torsion	base flipping torsion defined by the pseudodihedral angle (P1+P2)−P3−P4−P5
tGBs	torsion	simpler version of the glycosidic bond torsion defined by the dihedral angle A16 (O4′−C1−N9−C4)
tBFs	torsion	simpler version of the base flipping torsion defined by the dihedral angle by the axis A16(C4′−C3′) and the vectors T9(C3′)−A16(C4′) and A16(C3′−N1)
tBP1P2T	torsion	torsion of the entire segment of DNA defined by the pseudodihedral angle b1b2−p1−p2−t1t2
atan_dWCdHG	function	evaluation of the function atan2(dWC, dHG)

[a]The atoms involved in the centers of mass are highlighted in Figure 3A.

was selected to increase the difficulty for the framework, as it possesses overlap with ϕ and ψ during the transition.

Table 4 and Figure 3 describe all candidate CVs of the DNA system. The dihedral angles tGBs and tBF are based on refs 57 and 58 where they were successfully employed to obtain free energy landscapes of DNA base rolling and flipping. The distances dWC, dHG, dHB, dCC, and dNB and dihedrals tGB and tBF are taken from ref 59, where they were employed to obtain WC-to-HG paths and free energies. tGB is a more robust version of tGBs, which is able to correctly describe the rolling of the base even if the sugar rotates as well.[59] To observe the effects in the competition, we introduce yet another somewhat redundant CV; tBFs is a simplified version of tBF, which does not use centers of mass and is likely less reliable in the face of local deformations. dWC, dHG, dHB, and atan_dWCdHG were used in ref 56 to perform and analyze WC-to-HG path sampling simulations. tGBs, dCC, dHG, aBPT, and tBP1P2T are based on ref 63, where they were used to identify HG base pairs in a survey of X-ray structures. aBPT and tBP1P2T are the most large-scale CVs in the set, as they describe a deformation of the helix in the vicinity of the rolling base pair. dN1, dN2, and dN3 are added to the pool as CVs as they do not describe the transition individually but might complement some of the previously used CVs and factor into the competition. The remaining CVs, dNH-O_12_13 to dNH-N_4_21 in Table 4, have been included to fill the pool with less relevant CVs.

The final data sets consist of the CVs as features, i.e., inputs, and the Cartesian coordinates as labels, i.e., outputs, such that $\theta = \{CV_1^{(n)},...,CV_I^{(n)};\mathbf{x}^{(n)}\}_{n=1,...,N}$ is the training set over N frames with I different CVs in the candidate pool. For the alanine dipeptide case, all heavy backbone atom positions were selected as output. For the DNA case, the output coordinates included those of the transitioning base pair and its adjacent pairs and the backbone phosphorus atoms of the DNA strands, see Figure 4 for an illustration. All accepted trajectories of the TPS runs are combined and randomly rearranged, resulting in a randomized data set without any time-linearity.

RESULTS AND DISCUSSION

The results for the two molecular systems, alanine dipeptide and B-DNA, are discussed separately, and each is subdivided based on the two variants of the framework (current-time and

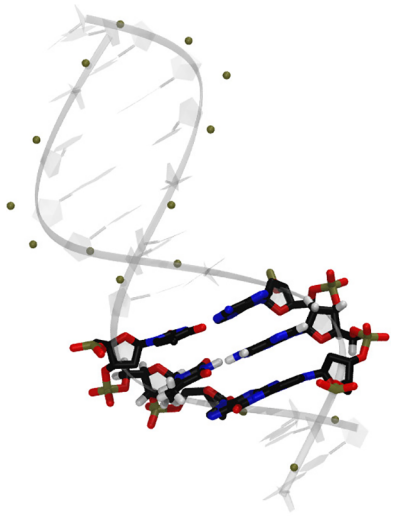


Figure 4. DNA atom selection used for NN output shown as balls and sticks. Based on the observed transition, the rearranging base pair atoms, the atoms of both adjacent base pairs, and the backbone phosphorus atoms were included.

https://dx.doi.org/10.1021/acs.jctc.0c00981
J. Chem. Theory Comput. 2021, 17, 2294−2306

lagged-time). In each subsection, we first discuss the architecture optimization, followed by the CV optimization.

The $C7_{ax}$-to-$C7_{eq}$ Transition in Alanine Dipeptide. *Current-Time Framework.* The framework was first applied to the alanine dipeptide data set, containing the CVs listed in Table 3. The entire pool of candidate CVs was selected as input for the initial NN architecture optimization step, as explained in the Methods section. The final optimal architecture of the neural network has two hidden layers, each consisting of 16 nodes with ReLU activation functions. To establish how "optimal" this final architecture is, multiple variations of it have been tested with 10 different runs using the same data set. The average results are recorded in Table 5 and show that the framework has at least identified a local optimum. A learning curve of an optimized individual is given in Figure 3 in the Supporting Information.

Table 5. Basic Variations of the Optimal Hidden Layer Architecture and Their Performance[a]

architecture	MAE loss (10^{-3})
optimal from framework	**7.0**
3 dense ReLU layers	7.6
4 dense ReLU layers	8.7
2 dense ReLU layers with 10% dropout	23.7
4 dense ReLU layers with 10% dropout	32.5
2 dense tanh layers	7.8
3 dense tanh layers	8.9
4 dense tanh layers	8.2

[a]The best performing architecture has been displayed in bold (lower MAE is better). No penalty has been applied regarding the number of inputs.

After this first step, the optimal NN hidden layer architecture was fixed, and the CV optimization was performed using the same settings, excluding the number of allowed input CVs, which was allowed to vary from 1 to 7. The penalty, introduced earlier, now results in a decreased number of selected CVs. The evolution in the CV selection converged quickly to three CVs: ϕ, ψ, and OO-distance dominate the pool after six generations (see Figure 9 in the Supporting Information).

In an attempt to quantify the importance of each of the selected CVs, we perform a sensitivity analysis by replacing each CV one at a time by noise,[64] such that all values of that CV in the set are uniformly distributed between 0 and 1, without retraining the network. The subsequent increase in loss is then a qualitative measure of how much that specific CV contributed to the overall recreation of the coordinates. Table 6 shows the losses of the trained network when one of the CVs has been replaced by noise. The highest increase in loss is observed when ϕ is replaced, followed by ψ, and finally dOO, suggesting that ϕ and ψ are more important than dOO, in accordance with previous results.[65]

Table 6. MAE Loss of the Trained Current-Time Network When One of the CVs Is Replaced by Noise[a]

CV	MAE loss (10^{-3})	Importance
none	4.4	
ϕ	28.1	5.3
ψ	23.7	4.4
dOO	12.1	1.7

[a]The importance is quantified as the increase of the loss relative to the original loss shown in the first row.

Lagged-Time Framework. The procedure was repeated with the inclusion of a time-lag in the output coordinates, arbitrarily chosen to be τ = 0.1 ps, limited by the shortest trajectory in the data set.

During optimization two distinct steps in the improvement of minimum loss were observed: the first corresponding to the domination of ψ and the second corresponding to the inclusion of ϕ (shown in Figure 11 in the Supporting Information). Following the same method as above, Table 7 shows the importance of the CVs, ranking ϕ above ψ. Some previous studies on alanine dipeptide (see, e.g., refs 16, 65, and 66) showed that the θ dihedral angle is also an essential CV. Instead, during the GA/NN optimization procedure, we observe an increase in the average loss when θ is selected (see, for example, the final five generations in Figures S8 and S9 in the Supporting Information). The most likely explanation for the discrepancy is the different force fields used in the various studies.

Table 7. MAE Loss of the Trained Lagged-Time Network When One of the CVs Is Replaced by Noise[a]

CV	MAE loss (10^{-3})	Importance
none	7.2	–
ϕ	41.9	4.8
ψ	23.2	2.2

[a]The importance is quantified as the increase of the loss relative to the original loss shown in the first row.

A Watson–Crick-to-Hoogsteen Transition in B-DNA. *Current–Time Framework.* In this second case study, the GA/NN framework is employed to obtain optimal CVs for the WC-to-HG transition of a base pair in a segment of DNA. Finding CVs for this transition in double-stranded DNA is a significantly more stringent test than the first case study because of the larger size of the system and the much higher complexity of the transition, involving H-bond breaking and formation, base rotation, base flipping, possible breathing modes in the neighboring base pairs, and possible bending and twisting of the double helix. In addition, we have added to the pool of candidates various CVs from previous studies for describing similar motions and which therefore show high degrees of correlation among each other (a table with the pairwise Pearson correlations between the CVs is shown in Figure 2 in the Supporting Information).

With all CVs selected as NN input, first the architecture of the hidden layers was optimized using the genetic algorithm GA 1. The optimization shows a small decrease in loss only for the first two generations of training, corresponding with fast convergence on the optimal architecture. The final architecture shows more complexity when compared to the optimal alanine dipeptide architecture and included four densely connected layers, consisting of 128, 64, 128, and 64 nodes in sequence, with activation functions ReLU, tanh, tanh, and SeLU, respectively. Layers 1 and 2 included batch normalization, while dropout was excluded for every layer.

Using this optimal architecture, the CV optimization was performed, resulting in selected CVs from three different categories: CVs that describe global movements within the molecule such as bending of the entire segment: aBPT and

https://dx.doi.org/10.1021/acs.jctc.0c00981
J. Chem. Theory Comput. 2021, 17, 2294–2306

tBP1BP2; CVs that describe the rotation and out-of-plane movement of the rotating base pairs: atan_dWCdHG and tBF; and distance CVs: dHB, dN1, dN2, dCC, dNB, and dN3. Of these, dHB, dN1, and dN2 correspond directly to the movement and rotation of the adenine, A16. The remaining CVs in this category, dCC, dNB, and dN3, correspond to the general conformation of the surrounding base pairs or the "pocket" in which the T9−A16 pair resides. Applying the same sensitivity analysis as before results in the CV importance ranking shown in Table 8.

Table 8. MAE Loss of the Trained Current-Time Network When One of the CVs Is Replaced by Noise[a]

CV	MAE loss (10^{-2})	Importance
none	3.5	−
dN2	11.3	2.2
dN1	9.8	1.8
tBF	9.8	1.8
dCC	8.6	1.5
dHB	7.1	1.0
atan_dWCdHG	6.3	0.8
dN3	5.8	0.7
dNB	5.1	0.5
aBPT	4.7	0.3
tBP1P2T	4.4	0.3

[a]The importance is quantified as the increase of the loss relative to the original loss shown in the first row.

The CVs found to describe this system well by Pérez de Alba Ortíz et al. are dWC, dHG, dHB, dCC, dNB, tGB, and tBF. Comparing the results, it can be seen that they generally agree, as tBF, dCC, dHB, and dNB are present in both. The framework did not select dWC, dHG, and tGB but instead chose atan_dWCdHG, dN1, and dN2, which cover similar movement. tGB is highly anticorrelated to the highly ranked dN2 and dN1 (with correlation coefficients of −0.62 and −0.77, respectively), indicating that the rolling of the adenine can be described by a combination of the two distances. dWC and dHG are closely related (0.64 and −0.56) to atan_dWCdHG, which is to be expected given the function's arguments. Moreover, in ref 56, atan_dWCdHG was successfully used as a reaction coordinate to extract a rate constants of the transition, which confirms its validity as an important CV. For a full overview of all the pairwise correlations, see SI Figure 2.

Lagged-Time Framework. Using the same settings, the procedure was repeated with the inclusion of a time-lag in the output, arbitrarily chosen to be 1 ps. The selected CVs can be divided into the three categories similar to the ones found in the current-time framework: CVs that describe structural changes far away from the rotating A16: dNH-N_4_21, dO-NH_12_13, and dO-NH_1_24; CVs concerning the rotation of A16: atan_dWCdHG and tGB; and distance CVs in the vicinity of A16: dWC, dCC, dHB, dN1, and dN3. The distances dN1 and dN2, which had the highest importance in the current-time selection, are now much less important, as they also take part in relatively fast breathing modes in the stacked base pairs, which is penalized by the time-lag. Instead, the rotation of the base pair is now covered by atan_dWCdHG, dWC, and tGB, while the out-of-plane movement can be described by a combination of dCC, dHB, dWC, and dN1. The remaining CVs show little importance for the transition.

As for the alanine dipeptide transition, the lagged-time framework shows a somewhat larger loss, and thus a worse prediction of the atom positions (MAE is 5.0×10^{-2}), than the current time version (MAE is 3.5×10^{-2}), which is not surprising because the lagged-time version filters fast-moving CVs, which leaves fewer input features for the network to construct an optimal fit. Nevertheless, the sensitivity analysis shows that the quality of the two solutions is very comparable; replacing (one of) the first CV(s) by noise in either framework solution results in a very similar loss (MAE: $11.3-11.4 \times 10^{-2}$). The advantage of the lag time is therefore that it provides a handle to enhance the selection of CVs that are less affected by fast irrelevant motions, which is helpful when the pool of candidates contains correlated CVs. In the alanine dipeptide case, the lagged-time framework removed the redundant dOO CV, and also in the DNA case, given a threshold importance, the number of important CVs is reduced in the lagged-time solution (see Table 9).

Table 9. MAE Loss of the Trained Lagged-Time Network When One of the CVs Is Replaced by Noise[a]

CV	MAE loss (10^{-2})	Importance
none	5.0	−
dWC	11.4	1.3
atan_dWCdHG	11.1	1.2
tGB	7.7	0.5
dCC	7.3	0.5
dN3	6.2	0.2
dHB	6.1	0.2
dN1	5.8	0.2
dO-NH_12_13	5.3	0.1
dO-NH_1_24	5.4	0.1
dNH-N_4_21	5.1	0.0

[a]The importance is quantified as the increase of the loss relative to the original loss shown in the first row.

Starting from Stable State Data, Alternating the Framework with Enhanced Sampling. So far, we have illustrated the GA/NN framework using trajectories obtained from TPS simulations as input data. TPS trajectories of the molecular transition are ideal for learning optimal descriptive CVs and thus for testing our framework; however, obtaining a representative TPS ensemble of trajectories can be very computationally demanding. Therefore, as an additional case in point, we apply the GA/NN framework on data from straightforward equilibrium simulations followed by alternating steps of enhanced sampling simulation and GA/NN CV extraction. This shows in the first place that the GA/NN framework does not depend on TPS trajectories to find optimal CVs and, second, that we are able to use the obtained CVs in an enhanced sampling simulation to compute accurate free energies. The molecular process under study is again the Watson−Crick-to-Hoogsteen transition in B-DNA.

We start from two 100 ns unbiased pre-equilibrated MD simulations, one sampling the WC state and the other sampling the HG state of the DNA segment. The simulation parameters are the same as in ref 56. From these trajectories, frames with an interval of 10 ps were taken and used as input for the framework. The architecture and settings were left unchanged with regard to the previous current-time DNA run.

https://dx.doi.org/10.1021/acs.jctc.0c00981
J. Chem. Theory Comput. 2021, 17, 2294−2306

The resulting CVs are listed in Table 10 in order of importance.

Table 10. CVs Resulting from the Stable State Data

CV	MAE loss (10^{-2})	Importance
tGB	0.12	1.96
tBF	0.06	0.42
dN1	0.05	0.34
dCC	0.05	0.29
dNH-O_12_13	0.05	0.16
dHB	0.05	0.14
dN-NH_12_13	0.05	0.13
aBPT	0.04	0.07
dNB	0.04	0.06
dN-NH_1_24	0.04	0.03
original	0.04	0.00

This initial set of CVs already contains several high-ranking CVs that overlap with the previous results obtained from TPS trajectory input. However, the set is contaminated with CVs that are obviously irrelevant for the description of the Watson−Crick-to-Hoogsteen transition but apparently enter because they enhance the NN prediction of DNA configurations sampled by other slow motions occurring during the equilibrium simulation. Second, since no configurations from the transition state or any other intermediate states along the transition were included in the input data, there is no guarantee that the obtained CVs are able to describe the rather nonlinear transition path. We thus proceed with an alternating sequence of enhanced sampling simulations, to obtain better representative configurations of the transition, and GA/NN application, to obtain a better set of CVs, in an iterative manner.

For the enhanced sampling step, we use restrained MD simulations along an adaptive path-CV to sample configurations along the transition outside the stable WC and HC states. Here, the path-CV is a function of a subset of CVs that is initially taken as a linear interpolation between the stable state values of these CVs. For the subset of CVs, we take the three highest scoring CVs shown in Table 10, i.e., tGB, tBF, and dN1. The progress parameter along the path-CV, s, equals 0.0 in the WC state (at tGB = 1.5 rad, tBF = −0.1 rad, dN1 = 4.1 Å) and 1.0 at HG (at tGB = −1.7 rad, tBF = 0.0 rad, dN1 = 6.3 Å). The path is implemented as a string of 20 nodes. Five parallel simulations (denoted walkers) are restrained at path progress values of s = 0.2, 0.4, 0.5, 0.6, and 0.8 with a stiff harmonic spring. The sampling is only restrained by the s-value but is free to move in a perpendicular direction away from the path, in the space spanned by the three CVs tGB, tBF, and dN1. We accumulate the average distance from the path to update the path nodes, with a frequency of 1 ps, so that the path-CV optimizes to the mean sampled density along the nonlinear reaction channel (see also refs 39 and 40 for details).

The sampling performed by these five walkers is shown in Figure S17 in the Supporting Information. We observe an outside pathway, i.e., with negative (−1.5 rad) tBF and large (>10 Å) dWC values for the intermediate states, similar to the pathway sampled by the TPS ensemble. The data corresponding to the last 9 ns of path-restrained MD sampling is reinserted into the GA/NN framework, resulting in the improved CV set shown in Table 11.

Table 11. CVs Resulting from the Path-Restrained MD Data

CV	MAE loss (10^{-2})	Importance
tBF	0.27	3.3
dHB	0.18	2.0
tBFs	0.12	1.0
tGB	0.10	0.60
dN1	0.10	0.59
dNB	0.08	0.23
tBP1P2T	0.07	0.19
dN3	0.07	0.15
aBPT	0.07	0.14
dO-NH_4_21	0.07	0.10
Original	0.06	0.00

From the second GA/NN analysis, we obtain the improved CVs: tGB, tBF, and dHB. It is worth noting the persistence of tGB and tBF in both the preliminary and the improved selection. Previously, this pair of CVs has been successfully used to compute free-energy surfaces for this transition.[58] We run again path-restrained MD with the improved set of CVs. The value of dHB is 3.0 Å in both the WC and HG states. This time, we employ a denser grid of 11 walkers, restrained from s = 0.0 to s = 1.0 with an interval of 0.1, which allows us to obtain the free energy profile along the optimized path. We again observe an outside pathway (see Figure S18 in the Supporting Information). The free-energy profile (see Figure 5) shows a barrier of 14 kcal/mol and a difference of 7 kcal/mol, resembling the results from refs 57−59.

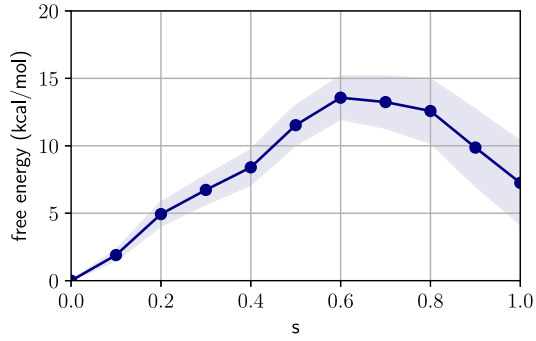


Figure 5. Free-energy profile of the WC-to-HG transition via the outside pathway calculated with path-restrained MD using the top CVs from the analysis reported in Table 11 (tGB, tBF, and dHB). The solid line and the shaded area represent the mean and standard deviation of 10 different profiles obtained from 10 blocks of 1 ns of simulation each.

Prediction of Atomic Coordinates. An interesting corollary of our framework is the generation of molecular configurations as output of the NNs from the set of optimal CVs. This was evaluated by visually inspecting the effect of changing a single CV on the configuration of the system. Figure 6 shows some results of applying this procedure on the best performing network of the current-time DNA run, by selecting the CV vector of a pre-existing configuration and changing only the value of a single CV from 0 to 1 while keeping the other CVs static. Here, the CV value is normalized, such that 0 refers to the WC state and 1 refers to the HG state.

Note that a significant change of a single CV, while keeping all other CVs fixed, requires the NN, trained on input from a rather concerted transition, to extrapolate to points in the CV

https://dx.doi.org/10.1021/acs.jctc.0c00981
J. Chem. Theory Comput. 2021, 17, 2294−2306

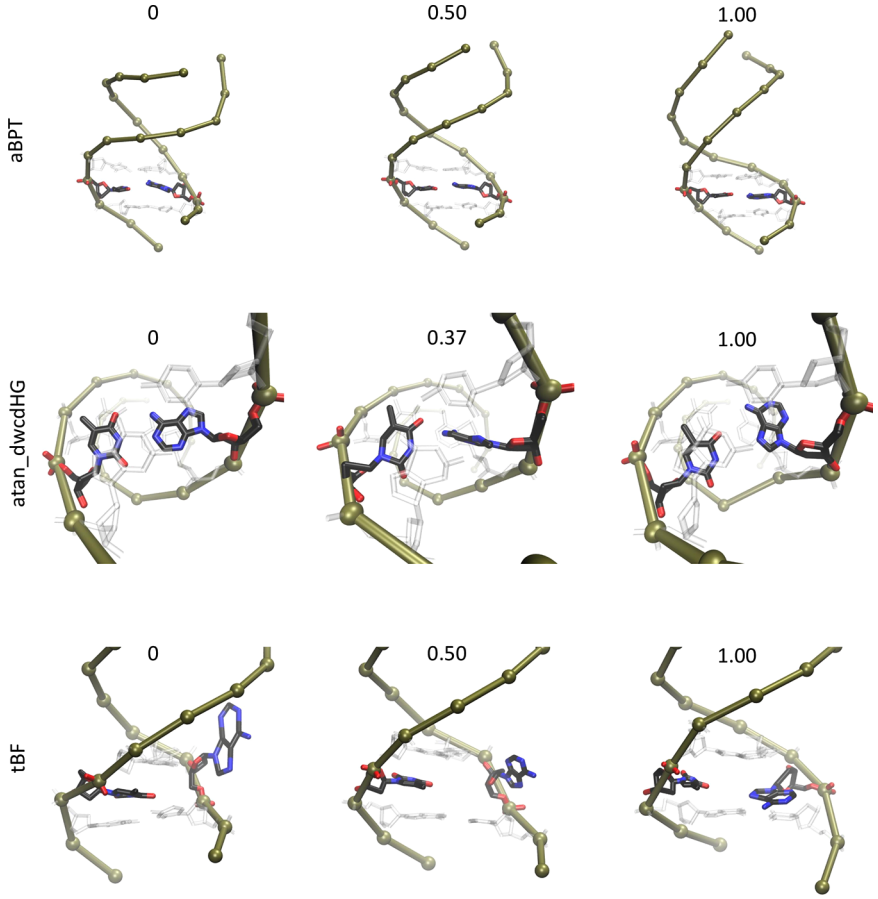


Figure 6. Configurations generated by the best performing NN of the current-time DNA run, by selecting the CV vector of a certain frame in the data set and changing a single CV within that vector to values ranging from 0 to 1 (see labels) while keeping the other CVs static. The first row displays the effect of modifying aBPT in a frame taken from the WC state as a degree of bending in the DNA strand through the phosphorus atoms. The second row shows the morphing of dWC to dHG of A16 in a frame taken from the WC state as a function of atan_dWCdHG. Lastly, the third row shows the out-of-plane movement of A16 as a result of modifying tBF in a frame taken from the transition state.

space that are far from the region sampled by the training set. Nevertheless, somewhat to our surprise, in all cases, the networks predict configurations that match rather well with our expectation for the structural change connected to the specific CV, albeit visibly with some nonphysical deformations of the molecular structure. From top to bottom in Figure 6, changing aBPT shows a bending in the entire double helix segment; modification of atan_dWCdHG results in (nonphysical) morphing of A16 from dWC to dHG; and modification of tBF induces the out-of-plane movement of A16. This suggests that this method may be used to generate molecular configurations in unexplored regions that can, after a short relaxation or equilibration, be used to accelerate sampling or to further improve the NN training in an iterative manner.

CONCLUSION

Taking inspiration from the pioneering work of Ma and Dinner,[19] we present a framework that uses supervised learning for automatic selection of CVs that optimally describe a molecular transition of interest. The framework assigns a fitness score to CVs based on their performance as input to a neural network that uses them to reconstruct the atom coordinates along the molecular transition. A genetic algorithm selects sets of CVs from a pool of candidates and evolves them toward the fittest set of CVs, which is the set that best allows the neural network to reproduce the molecular configurations of the transition. A second genetic algorithm optimizes the hyperparameters of the network, such as the number of hidden layers, the activation functions, etc. As the final CVs are a subset of the candidate pool, they are guaranteed to be as humanly understandable as the provided input.

The performance of the framework was illustrated at the hand of two case studies using input trajectory data from transition path sampling simulations. First, our method retrieved the established CVs (or equivalent) for the $C7_{ax}$-to-$C7_{eq}$ transition in alanine dipeptide, replicating the coordinates from the CVs with remarkable accuracy. Second, for the Watson−Crick-to-Hoogsteen transition of a base pair in B-DNA, our method retrieved local and global CVs previously found to be of importance in the literature: (1) atan_dWCdHG, the evaluation of atan2 (dWC, dHG) relating to the current hydrogen bond state of the base pair and found to be a suitable reaction coordinate for rate constant calculations;[56] (2) tGB and tBF, the base rolling and flipping torsions used in free-energy calculations of the transition;[57−59] (3) the C1−C1 distance (dCC) of the two base pairs; and (4) a degree of bending in the neighboring base pairs (aBPT and tBP1P2T) found in multiple X-ray structures of HG base pairs.[63]

A second variant of the framework, inspired by Wehmeyer et al.,[47] introduces a lag-time in the predicted atomic coordinates (output of the network) relative to the CV input to include

https://dx.doi.org/10.1021/acs.jctc.0c00981
J. Chem. Theory Comput. 2021, 17, 2294−2306

"the ability to predict the future" in the fitness score of the CVs. This lag-time acts as an extra filter that prefers CVs that describe the slow modes of the transition. For the alanine dipeptide case, adding a lag-time led to filtering of the redundant dOO CV, leaving only the well-known ϕ and ψ dihedral angles in the final set. Similarly in the DNA case, in which we provided several correlated CVs in the pool of candidates, the number of important CVs in the final set obtained with the lag-time was reduced, for example, by selection of a dihedral angle (tGB) instead of two distances (dN1 and dN2) that were selected by the current-time framework.

To show that the GA/NN framework does not depend on trajectories obtained from a converged TPS ensemble, we also showcased a practical application of the method starting from equilibrium trajectories of the stable reactant and product states. This already resulted in a useful set of CVs, which was subsequently used in an enhanced sampling (i.e., restrained MD) simulation to obtain an improved data set of configurations along the transition path. A second round of optimal CV retrieval and enhanced sampling was enough to obtain a CV set in agreement with the results obtained from the TPS data and, second, to obtain a converged, hysteresis-free, free energy profile of the WC-to-HG transition in agreement with previous results.

An interesting feature of the framework is that the trained neural network is able to generate atomic configurations when given an instance of the input CVs, which can be useful to setup simulations in unexplored microstates, using the framework for CV discovery in an iterative manner with MD simulations for data augmentation.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jctc.0c00981.

Details of neural network parameters, genetic algorithm parameters, pairwise Pearson correlation among CVs, and evolution of MAE loss functions during optimization (PDF)

■ AUTHOR INFORMATION

Corresponding Author

Bernd Ensing − *Van 't Hoff Institute for Molecular Sciences, AI4Science Laboratory, and Amsterdam Center for Multiscale Modeling, University of Amsterdam, 1098 XH Amsterdam, The Netherlands;* orcid.org/0000-0002-4913-3571; Email: b.ensing@uva.nl

Authors

Ferry Hooft − *Van 't Hoff Institute for Molecular Sciences, AI4Science Laboratory, and Amsterdam Center for Multiscale Modeling, University of Amsterdam, 1098 XH Amsterdam, The Netherlands*

Alberto Pérez de Alba Ortíz − *Van 't Hoff Institute for Molecular Sciences, AI4Science Laboratory, and Amsterdam Center for Multiscale Modeling, University of Amsterdam, 1098 XH Amsterdam, The Netherlands;* orcid.org/0000-0003-4776-5521

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jctc.0c00981

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

The authors acknowledge J. Vreede, P. G. Bolhuis, and D. W. H. Swenson for transition path sampling data of the B-DNA Watson−Crick-to-Hoogsteen base pairing transition and for valuable discussions about the system. A.P.d.A.O. gratefully acknowledges the Mexican National Council for Science and Technology (CONACYT) for financial support.

■ REFERENCES

(1) Torrie, G. M.; Valleau, J. P. Nonphysical sampling distributions in Monte Carlo free-energy estimation: Umbrella sampling. *J. Comput. Phys.* **1977**, *23*, 187−199.

(2) Carter, E. A.; Ciccotti, G.; Hynes, J. T.; Kapral, R. Constrained reaction coordinate dynamics for the simulation of rare events. *Chem. Phys. Lett.* **1989**, *156*, 472.

(3) Jarzynski, C. Nonequilibrium equality for free energy differences. *Phys. Rev. Lett.* **1997**, *78*, 2690.

(4) Laio, A.; Parrinello, M. Escaping free-energy minima. *Proc. Natl. Acad. Sci. U. S. A.* **2002**, *99*, 12562−12566.

(5) Bolhuis, P. G.; Chandler, D.; Dellago, C.; Geissler, P. L. Transition path sampling: Throwing ropes over rough mountain passes, in the dark. *Annu. Rev. Phys. Chem.* **2002**, *53*, 291.

(6) den Otter, W.; Briels, W. The reactive flux method applied to complex isomerization reactions: Using the unstable normal mode as a reaction coordinate. *J. Chem. Phys.* **1997**, *106*, 5494−5508.

(7) Komatsuzaki, T.; Berry, R. S. *Advances in Chemical Physics*; John Wiley & Sons, Ltd.: 2003; Chapter 2, pp 79−152, DOI: 10.1002/0471231509.ch2.

(8) Romo, T. D.; Clarage, J. B.; Sorensen, D. C.; Phillips, G. N. Automatic identification of discrete substates in proteins: Singular value decomposition analysis of timeaveraged crystallographic refinements. *Proteins: Struct., Funct., Genet.* **1995**, *22*, 311−321.

(9) Amadei, A.; Linssen, A. B.; Berendsen, H. J. Essential dynamics of proteins. *Proteins: Struct., Funct., Genet.* **1993**, *17*, 412−425.

(10) Elmaci, N.; Berry, R. S. Principal coordinate analysis on a protein model. *J. Chem. Phys.* **1999**, *110*, 10606−10622.

(11) Doruker, P.; Atilgan, A. R.; Bahar, I. Dynamics of proteins predicted by molecular dynamics simulations and analytical approaches: Application to α-amylase inhibitor. *Proteins: Struct., Funct., Genet.* **2000**, *40*, 512−524.

(12) Mendels, D.; Piccini, G.; Brotzakis, Z. F.; Yang, Y. I.; Parrinello, M. Folding a small protein using harmonic linear discriminant analysis. *J. Chem. Phys.* **2018**, *149*, 194113.

(13) Mendels, D.; Piccini, G.; Parrinello, M. Collective Variables from Local Fluctuations. *J. Phys. Chem. Lett.* **2018**, *9*, 2776−2781.

(14) Peters, B.; Trout, B. L. Obtaining reaction coordinates by likelihood maximization. *J. Chem. Phys.* **2006**, *125*, No. 054108.

(15) Lechner, W.; Rogal, J.; Juraszek, J.; Ensing, B.; Bolhuis, P. G. Nonlinear reaction coordinate analysis in the reweighted path ensemble. *J. Chem. Phys.* **2010**, *133*, 174110.

(16) Mori, Y.; Okazaki, K.-i.; Mori, T.; Kim, K.; Matubayasi, N. Learning reaction coordinates via cross-entropy minimization: Application to alanine dipeptide. *J. Chem. Phys.* **2020**, *153*, No. 054115.

(17) Fraser, A. S. Simulation of genetic systems by automatic digital computers I. Introduction. *Aust. J. Biol. Sci.* **1957**, *10*, 484−491.

(18) Rosenblatt, F. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review* **1958**, *65*, 386.

(19) Ma, A.; Dinner, A. R. Automatic method for identifying reaction coordinates in complex systems. *J. Phys. Chem. B* **2005**, *109*, 6769−6779.

(20) Pérez-Hernández, G.; Paul, F.; Giorgino, T.; De Fabritiis, G.; Noé, F. Identification of slow molecular order parameters for Markov model construction. *J. Chem. Phys.* **2013**, *139*, 015102.

(21) Mardt, A.; Pasquali, L.; Wu, H.; Noé, F. VAMPnets for deep learning of molecular kinetics. *Nat. Commun.* **2018**, *9*, 5.
(22) Hernández, C. X.; Wayment-Steele, H. K.; Sultan, M. M.; Husic, B. E.; Pande, V. S. Variational encoding of complex dynamics. *Phys. Rev. E: Stat. Phys., Plasmas, Fluids, Relat. Interdiscip. Top.* **2018**, *97*, No. 062412.
(23) Chen, W.; Sidky, H.; Ferguson, A. L. Capabilities and Limitations of Time-lagged Autoencoders for Slow Mode Discovery in Dynamical Systems. *J. Chem. Phys.* **2019**, *151*, No. 064123.
(24) Schöberl, M.; Zabaras, N.; Koutsourelakis, P. S. Predictive collective variable discovery with deep Bayesian models. *J. Chem. Phys.* **2019**, *150*, No. 024109.
(25) Chen, W.; Ferguson, A. L. Molecular enhanced sampling with autoencoders: On-the-fly collective variable discovery and accelerated free energy landscape exploration. *J. Comput. Chem.* **2018**, *39*, 2079−2102.
(26) Ribeiro, J. M. L.; Bravo, P.; Wang, Y.; Tiwary, P. Reweighted autoencoded variational Bayes for enhanced sampling (RAVE). *J. Chem. Phys.* **2018**, *149*, No. 072301.
(27) Zhang, J.; Chen, M. Unfolding Hidden Barriers by Active Enhanced Sampling. *Phys. Rev. Lett.* **2018**, *121*, No. 010601.
(28) Wu, H.; Mardt, A.; Pasquali, L.; Noe, F. In *Advances in Neural Information Processing Systems 31*; Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., Garnett, R., Eds.; Curran Associates, Inc.: 2018; pp 3975−3984.
(29) Tiwary, P.; Berne, B. J. Spectral gap optimization of order parameters for sampling complex molecular systems. *Proc. Natl. Acad. Sci. U. S. A.* **2016**, *113*, 2839−2844.
(30) Ceriotti, M.; Tribello, G. A.; Parrinello, M. Simplifying the representation of complex free-energy landscapes using sketch-map. *Proc. Natl. Acad. Sci. U. S. A.* **2011**, *108*, 13023−13028.
(31) Preto, J.; Clementi, C. Fast recovery of free energy landscapes via diffusion-map-directed molecular dynamics. *Phys. Chem. Chem. Phys.* **2014**, *16*, 19181−19191.
(32) Chiavazzo, E.; Covino, R.; Coifman, R. R.; Gear, C. W.; Georgiou, A. S.; Hummer, G.; Kevrekidis, I. G. Intrinsic map dynamics exploration for uncharted effective free-energy landscapes. *Proc. Natl. Acad. Sci. U. S. A.* **2017**, *114*, E5494−E5503.
(33) Rosso, L.; Mináry, P.; Zhu, Z.; Tuckerman, M. E. On the use of the adiabatic molecular dynamics technique in the calculation of free energy profiles. *J. Chem. Phys.* **2002**, *116*, 4389−4402.
(34) Maragliano, L.; Vanden-Eijnden, E. A temperature accelerated method for sampling free energy and determining reaction pathways in rare events simulations. *Chem. Phys. Lett.* **2006**, *426*, 168−175.
(35) Abrams, J. B.; Tuckerman, M. E. Efficient and Direct Generation of Multidimensional Free Energy Surfaces via Adiabatic Dynamics without Coordinate Transformations. *J. Phys. Chem. B* **2008**, *112*, 15742−15757.
(36) E, W.; Ren, W.; Vanden-Eijnden, E. Finite temperature string method for the study of rare events. *J. Phys. Chem. B* **2005**, *109*, 6688−6693.
(37) Vanden-Eijnden, E.; Venturoli, M. Revisiting the finite temperature string method for the calculation of reaction tubes and free energies. *J. Chem. Phys.* **2009**, *130*, 194103.
(38) Branduardi, D.; Gervasio, F. L.; Parrinello, M. From A to B in free energy space. *J. Chem. Phys.* **2007**, *126*, No. 054103.
(39) Díaz Leines, G.; Ensing, B. Path finding on high-dimensional free energy landscapes. *Phys. Rev. Lett.* **2012**, *109*, No. 020601.
(40) Pérez De Alba Ortíz, A.; Tiwari, A.; Puthenkalathil, R. C.; Ensing, B. Advances in enhanced sampling along adaptive paths of collective variables. *J. Chem. Phys.* **2018**, *149*, No. 072320.
(41) Rogal, J.; Schneider, E.; Tuckerman, M. E. Neural-Network-Based Path Collective Variables for Enhanced Sampling of Phase Transformations. *Phys. Rev. Lett.* **2019**, *123*, 245701.
(42) Leardi, R.; Boggia, R.; Terrile, M. Genetic algorithms as a strategy for feature selection. *J. Chemom.* **1992**, *6*, 267−281.
(43) Tang, J.; Alelyani, S.; Liu, H. *Data Classification*; CRC Press: 2014; Chapter Feature selection for classification: A review, pp 37−64.
(44) Vafaie, H.; De Jong, K. Genetic algorithms as a tool for feature selection in machine learning. *Proceedings Fourth International Conference on Tools with Artificial Intelligence TAI '92*; 1992; pp 200−203, DOI: 10.1109/TAI.1992.246402.
(45) Kudo, M.; Sklansky, J. Comparison of algorithms that select features for pattern classifiers. *Pattern Recognition* **2000**, *33*, 25−41.
(46) Van Rossum, G.; Drake, F. L. *Python 3 Reference Manual*; CreateSpace: Scotts Valley, CA, 2009.
(47) Wehmeyer, C.; Noé, F. Time-lagged autoencoders: Deep learning of slow collective variables for molecular kinetics. *J. Chem. Phys.* **2018**, *148*, 241703.
(48) *Keras*; 2015; https://github.com/fchollet/keras (accessed 2021-02-25).
(49) Hooft, F.; Pérez de Alba Ortíz, A.; Ensing, B. 2020. https://github.com/Ensing-Laboratory/FABULOUS (accessed 2021-02-25).
(50) Swenson, D. W. H.; Prinz, J.-H.; Noe, F.; Chodera, J. D.; Bolhuis, P. G. OpenPathSampling: A Python Framework for Path Sampling Simulations. 1. Basics. *J. Chem. Theory Comput.* **2019**, *15*, 813−836 PMID: 30336030.
(51) Swenson, D. W. H.; Prinz, J.-H.; Noe, F.; Chodera, J. D.; Bolhuis, P. G. OpenPathSampling: A Python Framework for Path Sampling Simulations. 2. Building and Customizing Path Ensembles and Sample Schemes. *J. Chem. Theory Comput.* **2019**, *15*, 837−856.
(52) Eastman, P.; Swails, J.; Chodera, J. D.; McGibbon, R. T.; Zhao, Y.; Beauchamp, K. A.; Wang, L.-P.; Simmonett, A. C.; Harrigan, M. P.; Stern, C. D.; Brooks, B. R.; Pande, V. S.; Wiewiora, R. P. OpenMM 7: Rapid development of high performance algorithms for molecular dynamics. *PLOS Computational Biology* **2017**, *13*, e1005659.
(53) Lindorff-Larsen, K.; Piana, S.; Palmo, K.; Maragakis, P.; Klepeis, J. L.; Dror, R. O.; Shaw, D. E. Improved side-chain torsion potentials for the Amber ff99SB protein force field. *Proteins: Struct., Funct., Genet.* **2010**, *78*, 1950−1958.
(54) Sivak, D. A.; Chodera, J. D.; Crooks, G. E. Using Nonequilibrium Fluctuation Theorems to Understand and Correct Errors in Equilibrium and Nonequilibrium Simulations of Discrete Langevin Dynamics. *Phys. Rev. X* **2013**, *3*, No. 011007.
(55) Essmann, U.; Perera, L.; Berkowitz, M. L.; Darden, T.; Lee, H.; Pedersen, L. G. A smooth particle mesh Ewald method. *J. Chem. Phys.* **1995**, *103*, 8577−8593.
(56) Vreede, J.; Pérez de Alba Ortíz, A.; Bolhuis, P. G.; Swenson, D. W. H. Atomistic insight into the kinetic pathways for WatsonCrick to Hoogsteen transitions in DNA. *Nucleic Acids Res.* **2019**, *47*, 11069−11076.
(57) Nikolova, E. N.; Kim, E.; Wise, A. A.; O'Brien, P. J.; Andricioaei, I.; Al-Hashimi, H. M. Transient Hoogsteen base pairs in canonical duplex DNA. *Nature* **2011**, *470*, 498−502.
(58) Yang, C.; Kim, E.; Pak, Y. Free energy landscape and transition pathways from Watson−Crick to Hoogsteen base pairing in free duplex DNA. *Nucleic Acids Res.* **2015**, *43*, 7769−7778.
(59) Pérez de Alba Ortíz, A.; Vreede, J.; Ensing, B. In *Biomolecular Simulations*; Bonomi, M., Camilloni, C., Eds.; Springer: 2019; Chapter 11, DOI: 10.1007/978-1-4939-9608-7_11.
(60) Ivani, I.; Dans, P. D.; Noy, A.; Pérez, A.; Faustino, I.; Walther, J.; Andrio, P.; Goñi, R.; Balaceanu, A.; Portella, G.; et al. Parmbsc1: a refined force field for DNA simulations. *Nat. Methods* **2016**, *13*, 55−58.
(61) Tribello, G. A.; Bonomi, M.; Branduardi, D.; Camilloni, C.; Bussi, G. PLUMED 2: New feathers for an old bird. *Comput. Phys. Commun.* **2014**, *185*, 604−613.
(62) Bonomi, M.; Bussi, G.; Camilloni, C.; Tribello, G. A.; Banas, P.; Barducci, A.; Bernetti, M.; Bolhuis, P. G.; Bottaro, S.; Branduardi, D.; et al. Promoting transparency and reproducibility in enhanced molecular simulations. *Nat. Methods* **2019**, *16*, 670−673.
(63) Zhou, H.; Hintze, B. J.; Kimsey, I. J.; Sathyamoorthy, B.; Yang, S.; Richardson, J. S.; Al-Hashimi, H. M. New insights into Hoogsteen base pairs in DNA duplexes from a structure-based survey. *Nucleic Acids Res.* **2015**, *43*, 3420−3433.

(64) Jung, H.; Covino, R.; Hummer, G. Artificial Intelligence Assists Discovery of Reaction Coordinates and Mechanisms from Molecular Dynamics Simulations. 2019, arXiv:1901.04595, arXiv e-prints. https://arxiv.org/abs/1901.04595 (accessed 2021-02-25).
(65) Bolhuis, P. G.; Dellago, C.; Chandler, D. Reaction coordinates of biomolecular isomerization. *Proc. Natl. Acad. Sci. U. S. A.* **2000**, *97*, 5877−5882.
(66) Ferguson, A. L.; Panagiotopoulos, A. Z.; Debenedetti, P. G.; Kevrekidis, I. G. Integrating diffusion maps with umbrella sampling: Application to alanine dipeptide. *J. Chem. Phys.* **2011**, *134*, 135103.

https://dx.doi.org/10.1021/acs.jctc.0c00981
J. Chem. Theory Comput. 2021, 17, 2294−2306