

University of Warsaw
Faculty of Physics

Maciej Dziubiski

TODO

Doctoral Thesis
TODO: na kierunku Fizyka
TODO: w zakresie Biofizyki

Praca wykonana pod kierunkiem
Prof. dr hab. Bogdana Lesynga
Zakład Biofizyki, Instytut Fizyki Doświadczalnej,
Wydział Fizyki, Uniwersytet Warszawski

Warsaw, September 2015

Chapter 1

Towards the identification of molecular cogs

Computer simulations of molecular systems allow determination of microscopic interactions between individual atoms or groups of atoms, as well as studies of intramolecular motions. Nevertheless, description of structural transformations at the mesoscopic level and identification of causal relations associated with these transformations, is very difficult. Structural and functional properties are related to free energy changes. Therefore, in order to better understand structural and functional properties of molecular systems, it is required to deepen our knowledge of free energy contributions arising from molecular subsystems in the course of structural transformations.

The method presented in this work quantifies the energetic contribution of each pair of atoms to the total free energy change along a given collective variable. Next, with the help of a genetic clustering algorithm, the method proposes a division of the system into two groups of atoms referred to as *molecular cogs*. Atoms which cooperate to push the system forward along a collective variable are referred to as *forward cogs*, and those which work in the opposite direction as *reverse cogs*.

The procedure was tested on several small molecules for which the genetic clustering algorithm successfully found optimal partitionings into molecular cogs. The primary result of the method is a plot depicting the energetic contributions of the identified molecular cogs to the total Potential of Mean Force (PMF) change. Case-studies presented in this work should help better understand the implications of our approach, and were intended to pave the way to a future, publicly available

implementation.

INTRODUCTION

Molecular dynamics (MD) simulation methods and advanced algorithms for calculating free energy bring us closer to predicting the physical properties of biomolecules [WR05; Chi14; Com+14]. However, computer simulations are not limited to interpreting experimental results. *In silico* one may also process MD data which can provide much more detailed information than that accessible in any experiment. The key to a deeper understanding of complex molecular systems is the extraction of valuable information from data produced in such simulations.

Biomolecules carry out their functions through conformational transitions between meta-stable states. Such phenomena can be simplified and described by a selected reaction coordinate which, in many cases, is a collective variable [CP07]. A valuable result of many simulation procedures (such as Umbrella Sampling [TV77], Thermodynamic Integration [Kir35], Adaptive Biasing Force [Com+14], and others [GRC12]), is a free energy profile, the so-called Potential of Mean Force (PMF) [Kir35]. In this study, the discussion is limited to one-dimensional PMFs, although it should be noted that the aforementioned methods can be formulated more generally and produce multi-dimensional free energy profiles. Of course, a one-dimensional PMF may be an oversimplification of what is going on in a complex system, but a meaningful collective variable usually leads to a free energy profile which makes the complicated transition more comprehensible [Per+12; Chi14]. However, what the PMF does not provide is the information about what drives transitions between meta-stable states.

Various attempts to understand and describe the internal mechanics of molecular systems have already been reported [AAT11; Bao09; Chi+13; LS07; Sac14; Aal+97; SG13]. We did not, however, find any general-purpose approach for studying a broader spectrum of cases, and in particular a methodology explaining the cause of a transition in terms of free energy contributions arising from certain parts of a molecule.

In this study we propose a new approach of analyzing the tendencies of a molecular system to undergo a selected structural transition. The main idea is to look at a shift along a collective variable as an effect of two opposing tendencies generated by interactions within the molecule. Our method indicates two groups

of atoms – referred to as *molecular cogs* – which, through cooperative interactions, are the source of these tendencies. For this purpose, we construct undirected graphs with weights between nodes expressed by energetic contributions to the free energy change arising from pair-wise interactions between atoms. These graphs are then partitioned into subsystems – corresponding to molecular cogs – using a genetic clustering algorithm. We present results for small, model systems which served as case-studies for the identification of molecular cogs and for testing if their functioning agrees with our intuitions.

METHODOLOGY

Decomposition of the Helmholtz free energy, A , was investigated in the past, most notably by Karplus and coworkers[BK95; BS95]. The aim was the determination of contributions to the free energy coming from components comprising the potential energy of the system. These potential energy components might come from different interaction types, or from cooperation between subsystems of the whole molecule. The strategy of expressing the potential energy of a system as a sum of terms, $U = \sum_i U_i$, and computing the contribution of each of these terms to the free energy is ineffective because additivity in the potential energy components does not imply additive contributions to the entropy[Dil97].

Two attempts of describing contributions to the free energy coming from parts of the system are worth mentioning. The first one was an approximate approach based on Free Energy Perturbation, in which higher order terms were neglected [BS95]. These terms are not, however, negligible, which severely limits the applicability of this method. An alternative route, employing Thermodynamic Integration, was also explored [BK95]. However, this approach is also limited, namely – values of the free energy contributions depend on the choice of the integration path.

We propose a different approach, in which pair-wise energetic contributions to the free energy are readily attained. Alas, as in the aforementioned methods of free energy decomposition, our current formulation struggles with a description of the entropic contributions, and at the present is not included in our method. The current implementation is primarily applicable to small molecules for which the role of entropy in structural transitions is negligible (see an example of a PMF in

the *Results* section).

Our analysis originates from the following formula [Car+89]:

$$A'(\xi^*) = \langle m_\xi \nabla U \cdot \mathbb{M}^{-1} \nabla \xi \rangle_{\xi^*} - \langle \mathbf{v} \cdot \nabla (m_\xi \nabla \xi) \mathbf{v} \rangle_{\xi^*}, \quad (1.1)$$

where A' is the derivative of the free energy with respect to the collective variable ξ , \mathbb{M}^{-1} is a diagonal matrix of inverse masses, and \mathbf{v} are velocities. The term m_ξ is defined by:

$$m_\xi := \left[\sum_i m_i^{-1} \left(\frac{\partial \xi}{\partial \mathbf{x}_i} \right)^2 \right]^{-1} \quad (1.2)$$

and can be interpreted as inertia of an effective point mass moving along the ξ coordinate.

On the right-hand side of Equation (1.1) we have, respectively, energetic and entropic contributions to the free energy change, both expressed as conditional averages, with the collective variable fixed at ξ^* . In the *Supporting Information* we briefly explain how these averages are estimated *via* constrained Molecular Dynamics (cMD) simulations, and highlight an important limitation of this method.

Note, that if we consider the potential energy as a sum of interaction components:

$$U = U^{ele} + U^{vdw} + U^{bond} + \dots = \sum_i^I U^I$$

the energetic contribution in Equation (1.1) maintains this additivity:

$$\langle m_\xi \nabla U \cdot \mathbb{M}^{-1} \nabla \xi \rangle_{\xi^*} = \sum_i^I \langle m_\xi \nabla U^I \cdot \mathbb{M}^{-1} \nabla \xi \rangle_{\xi^*}, \quad (1.3)$$

as was noted by Chipot et al. [CP07].

In our numerical experiments we used a force field with the following set of interaction types:

- *ele* (electrostatic interactions);
- *vdw* (van der Waals interactions);
- *bond* (2-body bonded interactions);
- *angl* (3-body angle interactions);
- *tors* (4-body torsional interactions).

We shall also consider:

- *nbd* (non-bonded interactions, the sum of *ele* and *vdw* interactions);
- *conf* (conformational interactions, composed of *bond*, *angl* and *tors* interactions);
- *total* (the sum of *nbd* and *conf* interactions).

Decomposition of the energetic contribution

Our purpose was to identify the molecular cogs, i.e. sets of atoms, which cooperate and push the whole system forward/backward along a reaction coordinate. We approached this problem by converting it into the task of finding clusters in a graph. Nodes in such a graph correspond to atoms, whereas edges represent cooperation of pairs of atoms. A natural measure of such cooperation can be introduced by taking into account the energetic contribution to the free energy in Equation (1.1).

Electrostatic, van der Waals and two-atom chemical bond interactions can be readily transformed into weights in the graph. For example, electrostatic interactions between atoms α and β lead to the following contribution:

$$c_{\alpha\beta}^{ele}(\xi^*) := \frac{1}{m_\alpha} \left\langle m_\xi \frac{\partial U^{ele}}{\partial \mathbf{x}_\alpha} \cdot \frac{\partial \xi}{\partial \mathbf{x}_\alpha} \right\rangle_{\xi^*} + \frac{1}{m_\beta} \left\langle m_\xi \frac{\partial U^{ele}}{\partial \mathbf{x}_\beta} \cdot \frac{\partial \xi}{\partial \mathbf{x}_\beta} \right\rangle_{\xi^*}. \quad (1.4)$$

Alas, n -body potential energy components cannot, in general, be decomposed into a sum of pair interactions. However, because the free energy contributions are of the form $\nabla U \cdot \mathbb{M}^{-1} \nabla \xi$, such decomposition is not required.

To clarify, let us consider a 3-body potential energy component, e.g. $U^{angl}(\mathbf{x}_\alpha, \mathbf{x}_\beta, \mathbf{x}_\gamma)$. The weight $c_{\alpha\beta}^{angl}$ of an edge joining nodes α and β represents the contribution coming from atoms α and β which interact *via* U^{angl} , while the γ atom is kept at a fixed position. By asserting that the sum of pair-wise contributions is equal to the total contribution arising from $U^{angl}(\mathbf{x}_\alpha, \mathbf{x}_\beta, \mathbf{x}_\gamma)$, i.e.:

$$\sum_{i>j} c_{ij}^{angl}(\xi^*) = \sum_{i=\alpha,\beta,\gamma} \frac{1}{m_i} \left\langle m_\xi \frac{\partial U^{angl}}{\partial \mathbf{x}_i} \cdot \frac{\partial \xi}{\partial \mathbf{x}_i} \right\rangle_{\xi^*},$$

we arrive at the following definition of the cooperation term:

$$c_{\alpha\beta}^{angl}(\xi^*) := \frac{1}{2} \sum_{i=\alpha,\beta} \frac{1}{m_i} \left\langle m_\xi \frac{\partial U^{angl}}{\partial \mathbf{x}_i} \cdot \frac{\partial \xi}{\partial \mathbf{x}_i} \right\rangle_{\xi^*},$$

where the $1/2$ means that the overall *angl* contribution is evenly distributed between the cooperation terms: $c_{\alpha\beta}^{angl}$, $c_{\beta\gamma}^{angl}$ and $c_{\alpha\gamma}^{angl}$.

For the general case of an n -body energy component, U^I , we propose the following definition of the cooperation term between atoms α and β :

$$c_{\alpha\beta}^I(\xi^*) := \frac{1}{n-1} \sum_{i=\alpha,\beta} \frac{1}{m_i} \left\langle m_\xi \frac{\partial U^I}{\partial \mathbf{x}_i} \cdot \frac{\partial \xi}{\partial \mathbf{x}_i} \right\rangle_{\xi^*}.$$

Note that we propose to distribute the cooperation evenly among all pairs of atoms involved in the interaction I . With this definition we constructed graphs for all interaction types, I .

A matrix, $\mathcal{C}^I(\xi^*)$, such that $[\mathcal{C}^I(\xi^*)]_{\alpha\beta} := c_{\alpha\beta}^I(\xi^*)$ constructed for a particular value of the collective variable ξ^* and the interaction type I is referred to as the *transient cooperation matrix for I* (see Figure 1.1). For convenience, the transient cooperation matrix for the *total* interaction type does not contain any superscript:

$$\mathcal{C}(\xi_i) := \sum_I \mathcal{C}^I(\xi_i).$$

Note that the sum $\sum_{\alpha>\beta} [\mathcal{C}(\xi_i)]_{\alpha\beta}$, is the overall energetic contribution to $A'(\xi_i)$ in Equation (1.1).

Comment on the entropic contribution

The second term on the right-hand side of Equation (1.1) does not explicitly depend on U . It is possible to propose pair-wise contributions of this quantity in order to construct matrices which in turn could be used as input for the genetic clustering procedure described below. However, calculating these entropic contributions would require computation of the Hessian of the collective variable (the matrix of second-order derivatives with respect to atom coordinates) in each step of the simulation, which would make the computational cost of the procedure prohibitively high. We speculate on a possible solution of this problem in the *Conclusions* section.

Clustering

Data clustering is the procedure of finding disjoint subsets (clusters) of objects that share a high affinity, and are dissimilar to objects from other clusters. To

carry out a clustering procedure we need a pair-wise affinity measure; thus, the data set is often represented by a weighted graph. Such a graph is encoded by an *affinity matrix* with elements corresponding to weights between nodes. In this study we were looking for groups of atoms whose cooperation decreased (increased) the free energy, thus pushing (pulling) the whole system forward (backward) along a reaction coordinate. Thus, we took the pair-wise energetic cooperation between atoms introduced earlier for the affinity measure.

We will refer to the clusters identified in a transient cooperation matrix as *transient molecular cogs*. The free-energy-reducing group is referred to as *forward cogs*, and the second group (which, conversely, increases the free energy gap) as *reverse cogs*. We, therefore, assume that a step taken by the molecule along a reaction coordinate is a consequence of a resultant tendency generated by the molecular cogs. But cross-cooperation between atoms from different groups can also occur, and we designate this unwanted effect 'gear grinding' (see the *Results* section).

Affinity Propagation algorithm

A considerable challenge associated with clustering is that different procedures produce different results. Therefore, it was crucial that the *right* clustering algorithm was chosen. We anticipated that the optimal algorithm should recover molecular cogs with as little gear grinding as possible, but we could not estimate how large this effect might be. To gain some insight into how molecular cogs might look like, we chose a clustering algorithm which is known to be successful in other applications.

Most clustering algorithms assume a non-negative affinity measure, whereas cooperation between atoms can be negative as well as positive. Among those algorithms which accepted negative values, Affinity Propagation (AP) was the procedure found to be the most promising[FD07]. The AP algorithm searches for *exemplars*, i.e. nodes within the graph around which non-exemplars are grouped, thus forming a cluster. It is an iterative procedure, where, in each step, "messages" between nodes are exchanged to designate exemplars and their followers.

In our case, the sign of cooperation is arbitrary i.e. it depends on the direction of the reaction coordinate, ξ . It was crucial to ensure that molecular cogs found

for a reaction coordinate $\xi' := -\xi$ were the same as those found for ξ . We resolved this problem by carrying out two clusterings with the AP method: one for $\mathcal{C}(\xi)$, and the second one for $-\mathcal{C}(\xi)$, in which the sign of all elements is changed, such as produced for ξ' . Description of the AP algorithm and our method for merging two clusterings can be found in *Supporting Information*.

One of the drawbacks of the AP method, similarly as in other clustering procedures, is that it requires several parameters, which influence the outcome. It appeared that small variations in parameters lead to markedly different results, and it was difficult to find a single set of parameters suitable for all cases (see the discussion in the *Supporting Information*). Nevertheless, the AP method gave us a first estimation of how molecular cogs look like, and it appears to be sufficiently good to initialize the genetic clustering procedure (see the *Genetic clustering* section).

Objective function

Molecular cogs identified by the AP algorithm were laden with low gear grinding, but that was not always a valuable finding. We noticed that in cases where no good partitioning was achievable, the AP method produced one-element clusters, which we considered to be artificial and incorrect.

The AP algorithm performs a clustering which maximizes the so-called *net similarity* [FD07]. This objective function, although helpful in many applications, does not carry any physical meaning. Note that cooperation, which we used to express affinity between atoms, translates into free energy differences, and thus into the tendency of the whole system to move along a collective variable. Molecular cogs should not only be laden with low gear grinding, but also cooperation generated by the cogs should cover the overall free energy change as much as possible.

To clarify, given a cooperation matrix \mathcal{C} , the sum of all negative (positive) elements is the overall tendency of the system to go forward (backward) along a reaction coordinate. It is desired for forward (reverse) cogs to cover as much of this overall tendency as possible. When there is only one atom in a cluster, there is no coverage, even though gear grinding might be small.

Let us denote by FC the set of atom indices, which were assigned to the forward

cogs, and analogously by RC the set of atoms assigned to the reverse cogs. Given a cooperation matrix, $\mathcal{C} = [c_{ij}]$, the following three quantities are useful in measuring the quality of molecular cogs:

- Forward Cogs Rate:

$$\text{FCR} := \frac{\sum_{i,j \in \text{FC}} c_{ij}}{\sum_{c_{ij} < 0} c_{ij}} \quad (1.5)$$

- Reverse Cogs Rate:

$$\text{RCR} := \frac{\sum_{i,j \in \text{RC}} c_{ij}}{\sum_{c_{ij} > 0} c_{ij}} \quad (1.6)$$

- Gear Grinding Rate:

$$\text{GGR} := \frac{\sum_{\substack{i \in \text{FC} \\ j \in \text{RC} \\ c_{ij} > 0}} c_{ij}}{\sum_{c_{ij} > 0} c_{ij}} + \frac{\sum_{\substack{i \in \text{FC} \\ j \in \text{RC} \\ c_{ij} < 0}} c_{ij}}{\sum_{c_{ij} < 0} c_{ij}} \quad (1.7)$$

We assumed that both positive and negative elements exist in the cooperation matrix, \mathcal{C} , i.e. that $\sum_{c_{ij} < 0} c_{ij} \neq 0$ and $\sum_{c_{ij} > 0} c_{ij} \neq 0$. In cases in which the denominator is null, we substitute the whole fraction by 0.

FCR is the ratio of the contribution captured by the forward cogs to the total forward propensity found in \mathcal{C} . RCR is analogous, and both these measures have a maximum value of 1. Magnitude of misplaced contributions between atoms from different clusters is indicated by GGR, which has a maximum value of 2.

We propose the following molecular cogs quality measure:

$$\text{SCORE}(\text{FC}, \text{RC}, \mathcal{C}) := 0.5(\text{FCR} + \text{RCR} - \text{GGR}), \quad (1.8)$$

and $\text{SCORE} := 0$ for cases when any of the cogs consists of a single atom. The above scoring function was used in our genetic clustering procedure (described in the following section). See Figure 1.2 for an example of a clustering which maximizes the SCORE.

SCORE equals 1 if the corresponding molecular cogs cover the whole propensity of the system to move along a reaction coordinate, and no gear grinding occurs. We observed that a SCORE less than 0.5 often indicates that the subdivision of the system into molecular cogs is noisy.

Note that a trivial partitioning in which all atoms are assigned to one group, e.g. FC, indicates that the system as a whole has a tendency to move forward. We

denote such trivial partitionings as “all FC” and “all RC”. The maximal value of SCORE is then 0.5, and although this is unfavorable, in many cases a trivial partitioning had the highest SCORE. In such cases, there was no convenient clustering into two groups, which means that the gear grinding was high.

Genetic clustering algorithm

Genetic algorithms are widely used in optimization problems when potential solutions are readily assessed and codified. They emulate the process of natural selection, promoting solutions – referred to as *specimens* – with higher scores. The codification of a solution is treated as its chromosome, which allows for *mutation* of a solution, and *crossover* with other chromosomes, thus “spawning” new specimens. At the end of each iteration of a genetic algorithm, there is a stage called *selection*, during which low-scoring solutions have a lesser chance of “survival”.

We used the genetic clustering algorithm to find molecular cogs with the highest SCORE, as defined in Equation (1.8) (see [Col98] for a good introduction to genetic clustering). The partitioning of a molecular system into molecular cogs was encoded by an array of numbers from the set $\{-1, 0, 1\}$, where the i th element of the array corresponded to the i th atom in a molecule. Values -1 and 1 translate into assignments to the forward and reverse cogs, respectively. An atom not belonging to any cluster was tagged by 0 ; this occurred when the atom did not cooperate with any other atom (see Figure 1.3).

To initialize the genetic clustering procedure, the AP algorithm was executed by applying five different methods of setting the diagonal elements in the affinity matrix (see *Supporting Information*). Next, it adds the “all FC” and “all RC” trivial solutions to the starting set. Following this, solutions are iteratively chosen, mutated and added to the set, until the starting population contained 200 candidate solutions.

Once the initial set is generated, the genetic procedure repeated the following steps:

1. Compose (randomly) 50 pairs of solutions to generate offspring using the crossover procedure.
2. Select (randomly) 20 solutions to generate offspring using the mutation pro-

cedure.

3. Calculate SCORE for the offspring and add it to the population.
4. Draw (randomly) 200 solutions from the population and discard the rest.
5. Choose the best scoring solution and check if there is any improvement in the SCORE. If there was none for 10 consecutive iterations, return the best solution. Otherwise, return to 1.

All random selections are done without repeats, so that a given solution with SCORE s is chosen with a probability proportional to e^{2s} . Although the number of parameters required in the genetic algorithm is daunting, changes to the majority of these parameters only influence the speed of arriving at the optimal solution (see the *Supporting Information* for a more detailed discussion of the parameter's influence on the outcome).

Figure 1.4 shows that the genetic clustering finds the optimal solution for a range of transient cooperation matrices for *nbd* (complete results can be found in the *Results* section). The optimal solutions were found by means of a brute-force search, i.e. by producing all possible partitionings and calculations of their SCOREs. It is worth noting that in all cases there was a singular solution with the highest SCORE.

Trapezoidal rule for integrating A

The free energy difference between ξ_X and ξ_Y can be expressed as an integral:

$$\Delta A = \int_{\xi_X}^{\xi_Y} A'(\xi^*) d\xi^*. \quad (1.9)$$

The free energy derivative, $A'(\xi^*)$, can be estimated at a given ξ^* from a cMD simulations *via* Equation (1.1). The integral in Equation (1.9) can then be calculated using the trapezoidal rule [Pre07]:

$$\widehat{\Delta A} \approx \frac{\Delta \xi}{2} \left\{ \widehat{A'(\xi_1)} + 2\widehat{A'(\xi_{i+1})} + \dots + 2\widehat{A'(\xi_{M-1})} + \widehat{A'(\xi_M)} \right\}, \quad (1.10)$$

where $\widehat{A'(\xi_i)}$ denotes an estimate of the free energy derivative at ξ_i . This requires independent cMD simulations at M values of the collective variable, $\{\xi_i\}_{i=1}^M$, with the grid size of $\Delta \xi$.

Each estimate $\widehat{A'(\xi_i)}$ has a corresponding variance, $\sigma^2[\widehat{A'(\xi_i)}]$, which in turn propagates to the variance of the integral estimate, i.e. $\sigma^2[\widehat{\Delta A}]$. In the following section we explain how we estimated the variances $\sigma^2[\widehat{A'(\xi_i)}]$. Because estimates $\widehat{A'(\xi_i)}$ are independent, the variance of the whole integral follows from Equation (1.10) straightforwardly:

$$\sigma^2[\widehat{\Delta A}] = \frac{\Delta \xi^2}{4} \{ \sigma^2[A'(\xi_1)] + 4\sigma^2[A'(\xi_2)] + \dots + 4\sigma^2[A'(\xi_{M-1})] + \sigma^2[A'(\xi_M)] \} \quad (1.11)$$

Details concerning estimation of the variances $\sigma^2[A'(\xi_1)]$ using a bootstrapping procedure can be found in the *Supporting Information*.

The same integration rule applies to all elements of the transient cooperation matrices. A matrix in which every element is a result of the above numerical integration is referred to as the *integrated cooperation matrix* or simply: *cooperation matrix*, and denoted by $\mathcal{C}(\xi_X \rightarrow \xi_Y)$. The sum $\sum_{\alpha > \beta} [\mathcal{C}(\xi_X \rightarrow \xi_Y)]_{\alpha\beta}$ is the energetic contribution to ΔA for the $\xi_X \rightarrow \xi_Y$ path. Molecular cogs found for this matrix are called *global molecular cogs*.

RESULTS

In the first part of this section we present detailed results for a small, 11-atom molecular model, to validate the concept of our theoretical approach. We attempted to indicate molecular cogs to verify whether the genetic algorithm finds the optimal clustering for the scoring function proposed in the *Objective function* section.

In the second part of the *Results* we show molecular cogs for the *nbd*, *ele* and *vdw* interactions for three other molecules. These case-studies allowed for testing of the transferability of the parameters used in the genetic clustering algorithm, but also uncovered certain subtleties characteristic to our approach.

The [NH3+]CC(I)I molecular model

We were interested in finding molecular cogs propelling a structural transition between two meta-stable states, separated by a high free energy barrier. For our first case-study we required a system with a fairly natural collective variable, in which all types of interactions are significant (*conf* as well as *nbd*). We used the

2,2-diiodoethan-1-aminium molecule, which in the SMILES format is encoded as: [NH3+]CC(I)I (we use the SMILES representation throughout this article because of its conciseness). The dihedral angle between atoms N1-C5-C6-I7 was our collective variable of choice (see Figure 1.5). We used the Generalized Amber Force Field (GAFF) to model interactions between atoms, with partial charges assigned using an empirical procedure, AM1-BCC (Table 1.1).

We focused on finding molecular cogs propelling the transition between the dihedral angles of $\xi_X := -172.5^\circ$ and $\xi_Y := -62.5^\circ$, which correspond to two PMF minima (Figure 1.5). The free energy barrier separating these minima is high, and interactions between the [NH3+] group and the iodine atoms provide strong *ele* and *vdw* contributions which influence this barrier.

Constrained MD simulations were carried out for fixed values of the ξ collective variable, each simulation with 10^5 1fs timesteps, at $T = 300\text{K}$. We chose 239 points, $\{\xi_i\}_{i=1}^{239}$, equally separated by $\Delta\xi = 0.5^\circ$, so that $\xi_1 = -179.0^\circ$ and $\xi_{239} = -60.0^\circ$. This was done in order to encompass the $[\xi_X, \xi_Y]$ interval; note that $\xi_{14} = \xi_X$ and $\xi_{234} = \xi_Y$.

The length of the block in the block bootstrap estimation of averages and their corresponding variances was set to 10^2 , which led to 10^3 blocks for each simulation data set. We chose the length of the block with the assumption that the auto-correlation after 10^2 steps is negligible in the case of our simple system.

The cMD procedure was implemented in the Python programming language (ver. 2.7.2), using the Open Babel[OBo+11] package (ver. 2.3.2) to model the molecule (see *Supporting Information*). All simulations for the [NH3+]CC(I)I model took about 30h on a desktop computer.

We focused on the free energy differences and energetic contributions for the $\xi_X \rightarrow \xi_Y$ path. For these end-points we obtain the free energy difference $\Delta A \approx -16.4 \frac{\text{kcal}}{\text{mol}}$, which is close to the energetic contribution, $\Delta E \approx -16.7 \frac{\text{kcal}}{\text{mol}}$. The residual $0.3 \frac{\text{kcal}}{\text{mol}}$ is the entropic contribution, which was negligible for our model.

Overview

Results for the *total*, *conf* and *nbd* interactions are juxtaposed in Table 1.2, and for 2-body interactions (*bond*, *ele*, *vdw*) in Table 1.3. The first row contains optimal

partitionings of the integrated cooperation matrices, i.e. global molecular cogs (see Figure 1.3 for explanation), along with a picture of the model colored according to the clustering. In the second row we placed the integrated cooperation matrices rearranged in accordance with the clusterings depicted in the previous row of the table (see Figure 1.2 for explanation). This representation visualizes the “density” of cooperativity within molecular cogs, and gear grinding. The third row shows SCOREs for transient molecular cogs and compares genetic clusterings with optimal, brute-force partitionings. In the next row we see an illustration presenting transient molecular cogs, which is a concise summary of how cooperation within the molecule changes with ξ . This plot is helpful in judging whether the global molecular cogs are similar to the transient molecular cogs, and in assessing the consistency of cooperation within the molecule. Finally, the last row contains a PMF-like contribution profile which we call the *energetic contribution profile*; this is the most important result of our analysis. The green line represents the total energetic contribution of a given interaction type, orange and blue lines show contributions of the reverse and forward cogs, respectively, and the purple line – the magnitude of gear grinding. Gear grinding is quantified as the sum of absolute values of all misplaced contributions, i.e. from pairs of atoms from different clusters. Forward and reverse cogs contributions are calculated as the sum of cooperation between atoms assigned to FC and RC, respectively.

Results for the $[\text{NH}_3^+]\text{CC}(\text{I})\text{I}$ molecular model

In this section we comment on the results presented in Tables 1.2 and 1.3. We explain the Tables row by row, highlighting important aspects of the analysis, starting with Table 1.2 and then proceeding with Table 1.3.

In the first row of Table 1.2, global molecular cogs for *total* and *conf* interactions are trivial (“all FC”), with very low SCOREs. Thus, no interesting cooperation was detected for these interaction types – the whole system has a general propensity for preferring the state around ξ_Y . On the other hand, global molecular cogs for the *nbd* interactions have a high SCORE. The clustering suggests, in particular, that non-bonded interactions between the nitrogen atom and the iodine atoms hinder the transformation (as expected). A more surprising implication was that

the hydrogen atoms from the [NH3+] group cooperate with the C6 atom.

Clustered cooperation matrices are placed in the second row of Table 1.2. For the *total* and *conf* interactions, for which the partitioning was trivial, the matrices kept their original form, whereas for the *nbd* interactions rows and columns were rearranged in accordance with the partitioning. Matrices for *total* and *conf* are almost indistinguishable, which suggests that the transition is mainly governed by the *conf* interactions. It is also clear that the cooperation matrix for *nbd* is less “dense”, with little gear grinding.

In the next row we see that the partitionings into transient molecular cogs for *nbd* have consistently higher SCOREs than those found for *total* and *conf*. There is a drop in SCOREs corresponding to the crossing of the free energy maximum at about -115° dihedral angle. SCOREs for *nbd* molecular cogs vary significantly for different values of the collective variable, and the division into forward and reverse cogs becomes slightly more difficult after crossing the free energy maximum.

The transient molecular cogs depicted in graphs in the fourth row of the table show that the transient molecular cogs for *total* and *conf* exhibit no interesting structure and, in most cases, are of the “all FC” or “all RC” type. Transient molecular cogs for *nbd* are fairly consistent with global molecular cogs. Interestingly, the previously noted *nbd* cooperation between hydrogen atoms H2, H3, H4 and the C6 atom persists throughout the transition.

In the last row of the table we placed the energetic contribution profiles. Because the global molecular cogs for the *total* and *conf* interactions were of the “all FC” form, the only contributions come from the forward cogs. For the *nbd* interactions, the plot shows a beautiful separation of contributions coming from the forward ($-4.2 \frac{\text{kcal}}{\text{mol}}$) and the reverse cogs ($3.7 \frac{\text{kcal}}{\text{mol}}$) for non-bonded interactions, and low gear grinding for the $\xi_X \rightarrow \xi_Y$ transition.

Let us now look at the results in Table 1.3. In all three cases global molecular cogs were non-trivial, although the SCORE for *bond* is significantly lower than for *ele* and *vdw*. The partitioning for *bond* suggests that there is an impediment arising from interactions of the C6, I7 and I8 atoms, whereas the rest of the system favors the state around ξ_Y . Moving on to *ele* and *vdw* interactions we see that their global molecular cogs share a common pattern, although the SCORE for

the latter is slightly lower. We can also see that the cooperation of the H2, H3, H4 and C6 atoms (indicated earlier for *nbd*) is caused by the *ele* interactions.

The next row of Table 1.3 shows graphical representations of the clustered cooperation matrices. Not surprisingly, the *bond* matrix is more sparse than any other matrix, however judging by the low SCORE for global molecular cogs, this was not sufficient to ensure a clear division into the forward and reverse cogs. Conversely, for *ele* and *vdw* there seems to be a higher degree of gear grinding, yet the SCOREs were higher. This is due to the fact that the cooperation within molecular cogs is much stronger than gear grinding between them.

SCOREs in the third row of the table show that the transient molecular cogs for the *bond* interactions were consistently low. We see the opposite for *ele*, and a completely different situation for the *vdw* transient molecular cogs. The case of the *vdw* cooperation is particularly interesting because it shows that the global molecular cogs can have a high SCORE despite the fact that most of the transient molecular cogs have SCOREs below 0.5.

In the fourth row of the table we see that the transient molecular cogs for *bond* are consistent with their global molecular cogs for dihedral angles in the $[-150^\circ, -100^\circ]$ interval (i.e. around the free energy maximum at $\xi \approx -115^\circ$) For the *ele* interactions we see a stable cooperation between the H1, H2, H3 and C6 atoms, occasionally aided by atoms: H9 and I8. The reason behind the shape of the *vdw* SCORE plot becomes clearer once we see that the cooperation for this interaction type has two stages – before and after crossing the free energy maximum. Transient molecular cogs for *vdw* on the left side of the free energy maximum are mainly trivial (“all RC”) with a SCORE of about 0.5. To the right of the maximum there is a change in cooperation; we see a steady partitioning into forward cogs composed of atoms: N1, I7, I8, H9, H10 and H11, and reverse cogs (atoms: H2, H3, H4 and C6).

In the last row of the table we see that the *bond* interactions lead to molecular cogs with high gear grinding, which is the cause of low SCOREs. The profile shows that these interactions lower the free energy gap in the $\xi_X \rightarrow \xi_Y$ transition by $-9.9 \frac{\text{kcal}}{\text{mol}}$. However, we can also see that the contribution from atoms C6, I7 and I8 alone increases this gap by $5 \frac{\text{kcal}}{\text{mol}}$. For the *ele* interactions the separation into

molecular cogs was clean (low gear grinding), with a $-4.3 \frac{\text{kcal}}{\text{mol}}$ contribution from the forward cogs, and a $2.8 \frac{\text{kcal}}{\text{mol}}$ contribution from the reverse cogs due to the transition. For *vdw*, the separation has also led to low gear grinding, however the overall contribution from the forward cogs is much smaller than the one from the reverse cogs. This is an important observation which occurs again in the next section, where we analyze global molecular cogs determined for other model systems.

Results for related molecules

To better understand our approach, it is valuable to identify molecular cogs for other systems, related to the $[\text{NH}_3^+]\text{CC}(\text{I})\text{I}$ molecule scrutinized above. It was also valuable from the perspective of testing the transferability of the parameters used in the genetic clustering algorithm (as shown in the full results in the *Supporting Information*). We took into account molecules which instead of the $[\text{NH}_3^+]$ group had the: NH_2 , CH_3 and CH_2Cl groups respectively, i.e. molecules: $\text{NCC}(\text{I})\text{I}$, $\text{CCC}(\text{I})$ and $\text{CClCC}(\text{I})\text{I}$. Partial charges assigned by the AM1-BCC procedure for these molecules are listed in Table 1.1. In this section we report and shortly discuss global molecular cogs for the *nbd*, *ele* and *vdw* interactions (see Table 1.4).

Because we analyzed closely related molecules, it was expected that molecular cogs for the *nbd* interactions would be comparable. All molecules presented here, except for $\text{CClCC}(\text{I})\text{I}$, have partial charges of the same sign for corresponding atoms, therefore global molecular cogs for *ele* look similar. The more interesting outcome was related to discrepancies in molecular cogs identified for the *vdw* interactions.

As noted earlier, in the case of the $[\text{NH}_3^+]\text{CC}(\text{I})\text{I}$ molecule the overall *vdw* contribution is almost entirely explained by cooperation within the reverse cogs. The forward cogs were identified merely because their contribution was non-positive. Nevertheless, this was a valuable insight – we have learned that the *vdw* steric effects are due to interactions of particular atoms: those which comprise the reverse cogs. On the other hand, the *vdw* molecular cogs for the $\text{NCC}(\text{I})\text{I}$ molecule are trivial (“all RC”) and carry no such information. No non-trivial partitioning with a higher SCORE was found because there were no legitimate forward cogs, even as ineffective as those discovered in $[\text{NH}_3^+]\text{CC}(\text{I})\text{I}$. This then suggests that we gained

a simplified picture of *vdw* cooperation in the $[\text{NH}_3^+]\text{CC}(\text{I})\text{I}$ molecule simply because we were fortunate enough to have analyzed a system in which there had been at least a minimal contribution from the forward cogs. This is a consequence of the underlying assumption that a molecule’s tendency to undergo a transition is a result of two opposing cooperations. Perhaps we should approach the problem differently for instances in which the molecule as a whole has the propensity to move forward/backward (as we also saw for the *total* and *conf* interactions).

Molecular cogs for *vdw* interactions for the $\text{CCC}(\text{I})\text{I}$ molecule are similar to those of $[\text{NH}_3^+]\text{CC}(\text{I})\text{I}$. This might seem natural; the $[\text{NH}_3^+]$ and CH_3 groups share common properties. However, the energetic contribution profile for the *vdw* interactions (see *Supporting Information*) reveals that the forward cogs in $\text{CCC}(\text{I})\text{I}$ have a contribution comparable to that of the reverse cogs. Although *vdw* molecular cogs for $\text{CCC}(\text{I})\text{I}$ and $[\text{NH}_3^+]\text{CC}(\text{I})\text{I}$ are identical, the underlying mechanism is different.

The $\text{CClCC}(\text{I})\text{I}$ model was designed to lower the *ele* barrier. However, the resulting molecular cogs for *ele* became trivial and, as in the case of *vdw* for $\text{NCC}(\text{I})\text{I}$, we do not know which atoms are the main source of this effect. This again suggests that perhaps an alternative method of finding molecular cogs should be considered. In cases of trivial partitionings, such method should extract information about parts of the molecule which are the main source of the free energy difference. However, we leave investigation of properties and characteristics of alternative scoring functions for future studies.

DISCUSSION

Note that in all cases the genetic clustering algorithm gave clusterings with the best possible SCOREs. However, the efficiency of the genetic algorithm depends on the starting point, and without the help of the AP-generated initial population it took, on average, about 30 times longer, and the best result was not always achieved. To generate these initial solutions we adapted the AP clustering procedure (see *Supporting Information*).

The reason why conformational interactions lead to low-quality molecular cogs is that these are short-ranged interactions and our test molecular system is small.

Partitioning of a graph into clusters is laden with a cost which depends on the weights of cross-cluster edges. But, as can be seen in Table 1.2, the integrated cooperation matrices for non-bonded interactions are more sparse than for conformational interactions. Consequently, the cost of partitioning is greater for more “dense” matrices. But this cost decreases with increasing dimensionality of a matrix. It is therefore possible that, for more complex systems, molecular cogs for conformational interactions will have a higher SCORE.

The non-bonded interactions in the GAFF force field for atoms separated by three bonds or fewer are zero. This, and the fact that our model system is small, led to a sparse *nbd* cooperation matrix. Note, however, that a larger system may yield matrices that would be more “dense” for non-bonded interactions than for short-ranged conformational interactions. Whether a partitioning of these matrices would result in higher gear grinding remains an unanswered question.

We also anticipate that the entropic contribution in Equation (1.1) should play a more notable role for larger systems. Nonetheless, we have not yet considered clustering a graph with edges weighted by the entropic contributions of pairs of atoms. The reasons are twofold: practical (we want to avoid calculating Hessians of collective variables) and conceptual (the entropic cooperation of an atom with itself is non-zero). Entropic contributions need to be considered, but this should be the subject of future studies.

In the second part of the *Results* section we identified and analyzed molecular cogs for three additional molecules, related to the [NH3+]CC(I)I model. We discovered that trivial molecular cogs, which carry less information than any other partitioning, may occur whenever there is a lack of competition of cooperation within the molecule. This is a consequence of the proposed approach of dividing the system into two competing subsystems. In many cases in which global molecular cogs were trivial, we simply did not gain any insight into what is propelling the transformation. However, it seems that in such cases the analysis should be aimed at finding parts of a molecule that play the dominant role in its mobility.

Comparison with qualitative expectations

Results for the $[\text{NH}_3^+]\text{CC}(\text{I})\text{I}$ model, especially for the *ele* interactions, were useful in checking our intuitions against what was shown by the analysis. In this section we shortly discuss several sanity checks related to results for the *ele* interactions that were helpful in verifying whether our implementation had any critical errors, and whether the proposed approach leads to reasonable assessments.

From the plot depicting transient molecular cogs for *ele* (Table 1.3) we see that atoms N1 and I7 consistently hinder the transformation. These two atoms have negative partial charges (-0.85 and -0.08 , respectively; see Table 1.1), and therefore repel each other throughout the transition. But the I8 atom (partial charge of -0.08) behaves differently, i.e. for ξ lower than -115° (free energy maximum) it impedes the process by repelling the N1 atom, and aids it for ξ larger than -115° . This effect was correctly captured by the above analysis, because the I8 atom finds itself in the reverse cogs in the first part of the transition, and in the forward cogs in the second part (as shown by the graph in the sixth row of Table 1.3). From the integrated contribution matrix for *ele* (second row in Table 1.3) we see, however, that the cumulative contribution from the N1-I8 atoms is positive, which results in assigning them to one cluster.

The analysis revealed that atoms: H2, H3, H4 and C6 were consistently cooperating electrostatically, lowering the free energy barrier. This conclusion was more unexpected than the one concerning atoms: N1, I7 and I8, but was also compatible with our intuitions, taking into account the partial charges in Table 1.1.

Our method can suggest a qualitative interpretation of the cause behind a structural transition. However, it should be noted the most valuable information gained from this type of analysis is the quantitative description of the molecular cogs *via* the energetic contribution profile.

CONCLUSIONS

The aim of this article was to introduce a new methodology of identifying molecular cogs – parts of a molecule that propel structural transitions in forward/backward directions along a collective variable. The current framework allowed us to track energetic contributions to the free energy, leaving the problem of including entropic

terms for future developments. Results show that with the use of the genetic clustering algorithm we can successfully divide small molecules and identify forward and reverse molecular cogs associated with non-bonded interactions.

In particular, we proposed the approach of defining free energy contributions originating from pairs of atoms, and a method of dividing a molecule into molecular cogs. We showed that the proposed genetic clustering algorithm efficiently finds the optimal cogs, leading to high-quality partitionings for non-bonded interactions. However, we also found that conformational interactions lead to low-scoring molecular cogs, and that the system as a whole favored one meta-stable state over the other.

Currently, our method is based on cMD simulations for computing conditional averages (Equation (1.1)). Unfortunately, this solution has a critical drawback (see *Supporting Information*), but also, in order to determine the entropic contributions, requires computing second-order derivatives of the collective variable. To resolve this problem we should facilitate the Adaptive Biasing Force (ABF) scheme for calculating the PMF [Com+14]. Specifically, our future work is aimed at reformulating the procedure as a plugin to the NAMD package (in which the ABF has already been implemented), to include entropic contributions and to assure scalable performance. Once this is done, numerous new opportunities will become available, some of which we mention below.

We constructed graphs in which nodes corresponded to atoms. Note, however, that they may also be assigned to amino acids (or more sizable objects) to construct graphs, which would lead to a more mosaic clustering. It could also be valuable to represent a whole ligand as a single object in a protein-ligand complex. One could also consider the role of functional groups comprising a ligand and amino acids in a binding pocket; it might then be helpful to represent the rest of the protein as a single node in the graph. Another example is a possible treatment of solvent molecules, for example: a particular group of interesting water molecules could be transformed into separate nodes in the graph, while others reduced to a single node.

Our method may prove helpful in understanding why wild-type proteins perform better than mutants, or in explaining why certain drugs perform better than others,

despite their structural similarity. It would also be interesting to consider a multi-stage induced fit docking, and to study cooperativity between amino acids and the ligand along a collective variable.

This research study demonstrates how to extract pair-wise energetic contributions to the free energy change along a reaction coordinate. In its present form¹, we could only use our method to analyze small, model molecules. However, we were able to verify the efficacy of the genetic clustering algorithm, and to learn what can be expected from the new notion of "molecular cogs". Our future aim is to analyze more complex systems, and to develop a publicly available implementation of our method for practical applications.

¹We made the source code available at: <http://github.com/ponadto/molecular-cogs>

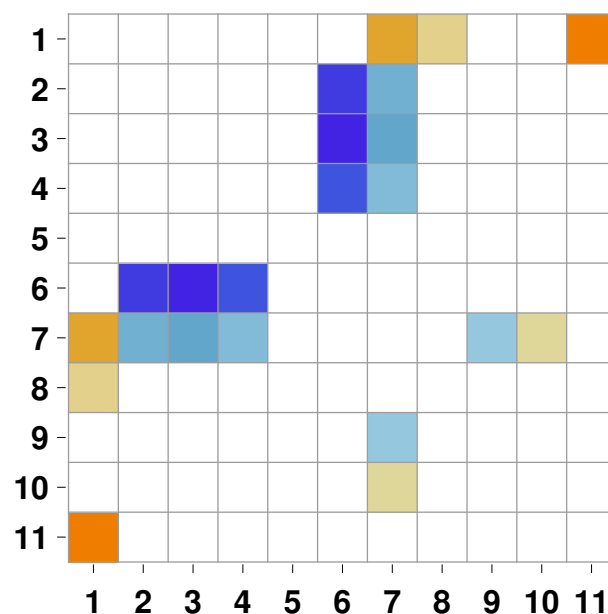


Figure 1.1: **Graphical representation of a transient cooperation matrix for *nbd* interactions.** Warm colors denote positive contributions, whereas blue – negative. White squares correspond to nil values.

[NH3+]CC(I)I		NCC(I)I		CCC(I)I		CC1CC(I)I	
atom	partial charge	atom	partial charge	atom	partial charge	atom	partial charge
N1	-0.85	N1	-0.92	C1	-0.10	C1	0.03
H2	0.47	H2	0.36	H2	0.04	Cl2	-0.19
H3	0.47	H3	0.36	H3	0.04	H3	0.08
H4	0.47			H4	0.04	H4	0.08
C5	0.13	C4	0.17	C5	-0.07	C5	-0.08
C6	0.08	C5	0.20	C6	0.21	C6	0.20
I7	-0.08	I6	-0.19	I7	-0.19	I7	-0.19
I8	-0.08	I7	-0.19	I8	-0.19	I8	-0.19
H9	0.13	H8	0.05	H9	0.06	H9	0.08
H10	0.13	H9	0.05	H10	0.06	H10	0.08
H11	0.13	H10	0.12	H11	0.10	H11	0.12

Table 1.1: **Partial charges assigned by the AM1-BCC procedure.**

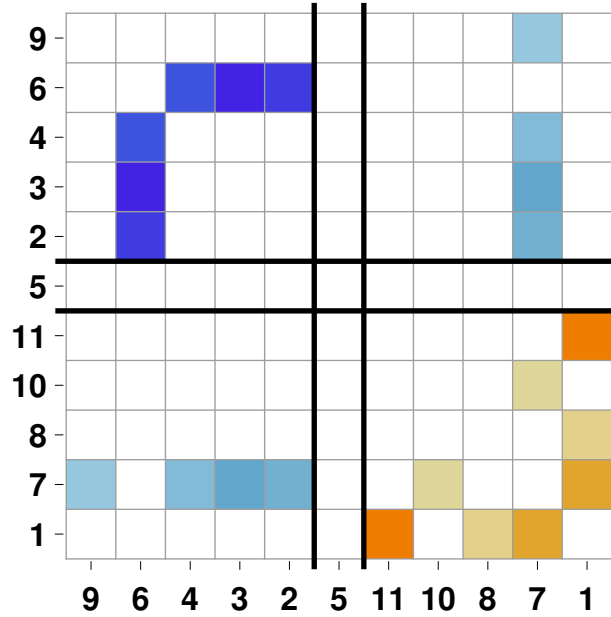


Figure 1.2: **Graphical representation of a result of a clustering procedure.** Rows and columns of the cooperation matrix in Figure 1.1 were reordered according to their assignments to: forward cogs, non-interacting atoms, and reverse cogs. In the upper left corner of this transformed matrix we have a block (square submatrix) with negative contributions, which come from the atoms 2,3,4 and 6 (from the forward cogs). In the lower right corner we have a block formed by pair contributions of the atoms in the reverse cogs. The gear grinding is small and results from contributions between the atom 7 and the atoms 2,3,4,9.

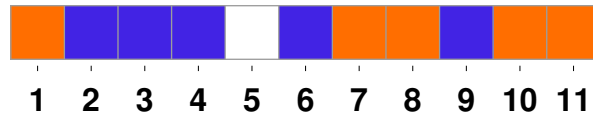


Figure 1.3: **Graphical representation of an exemplary partitioning:** $\{1, -1, -1, -1, 0, -1, 1, 1, -1, 1, 1\}$. Orange squares correspond to 1 (reverse cogs), blue to -1 (forward cogs), and white to 0 (non-interacting). The cooperation matrix in Figure 1.1 was reordered according to this assignment, yielding the transformed cooperation matrix in Figure 1.2.

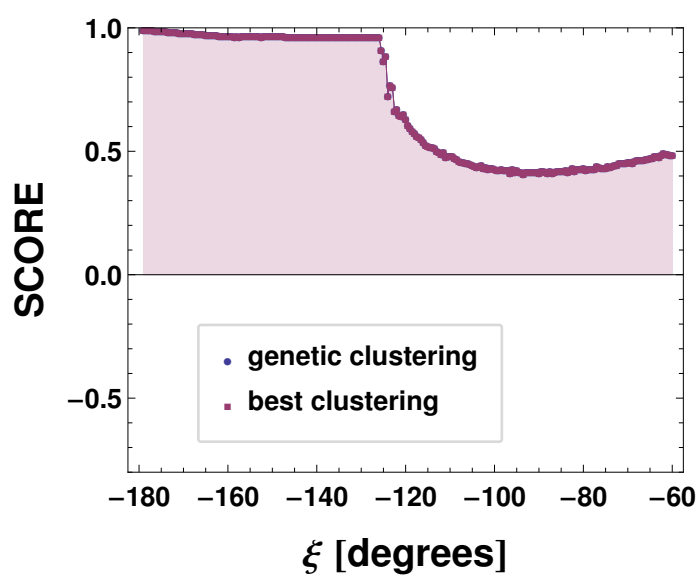


Figure 1.4: **SCORE** plot for *nbd* interactions. The plot shows that solutions found by the genetic clustering algorithm overlap perfectly with the highest scoring assignments to the transient molecular cogs.

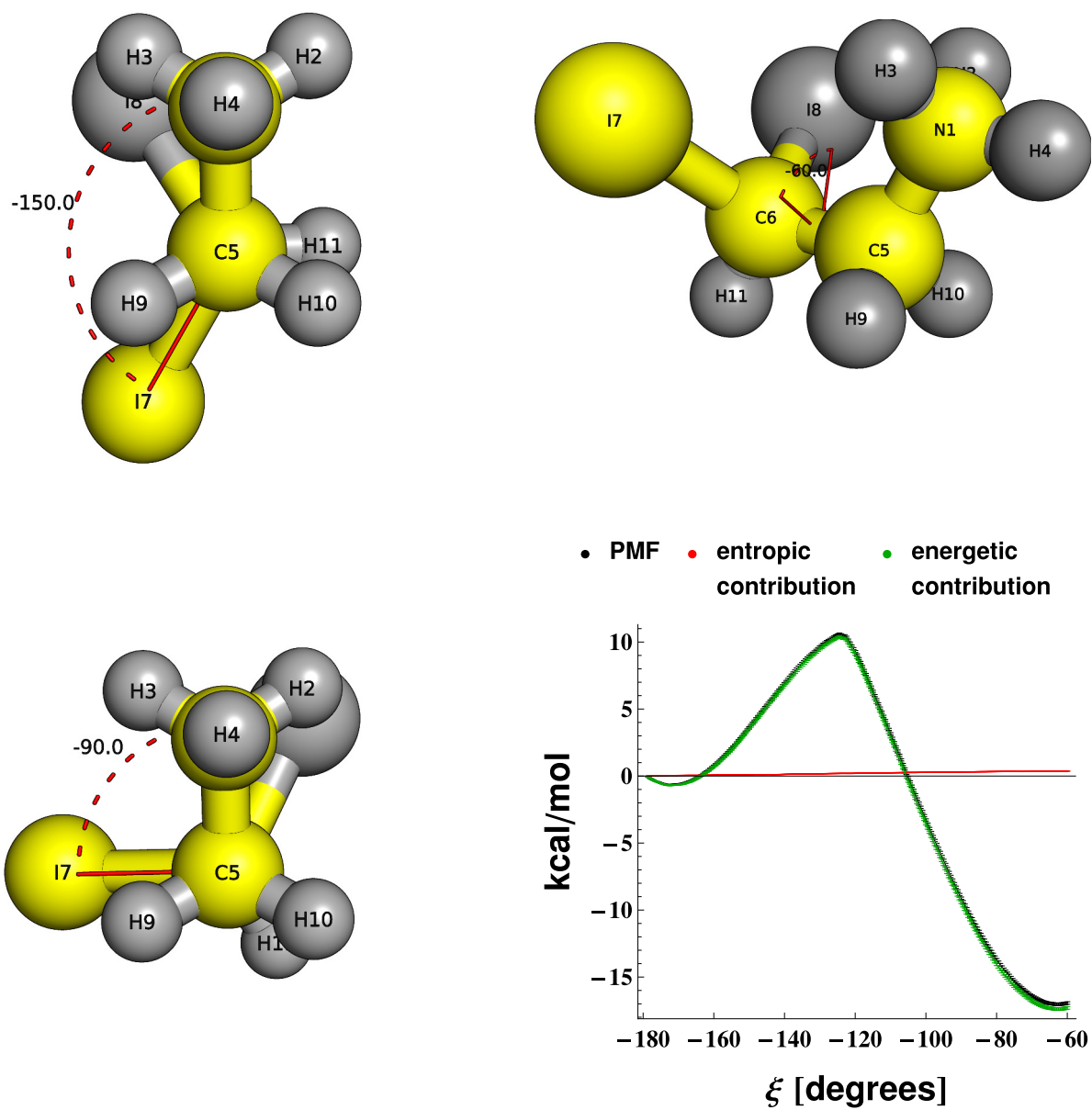


Figure 1.5: $[\text{NH}_3^+]\text{CC}(\text{I})\text{I}$ **molecule summary**. Three configurations of the molecule are shown, and the PMF with energetic and entropic contributions (for $T = 300$ K). Atoms N1-C5-C6-I7 (yellow) were used to define the dihedral angle, ξ .

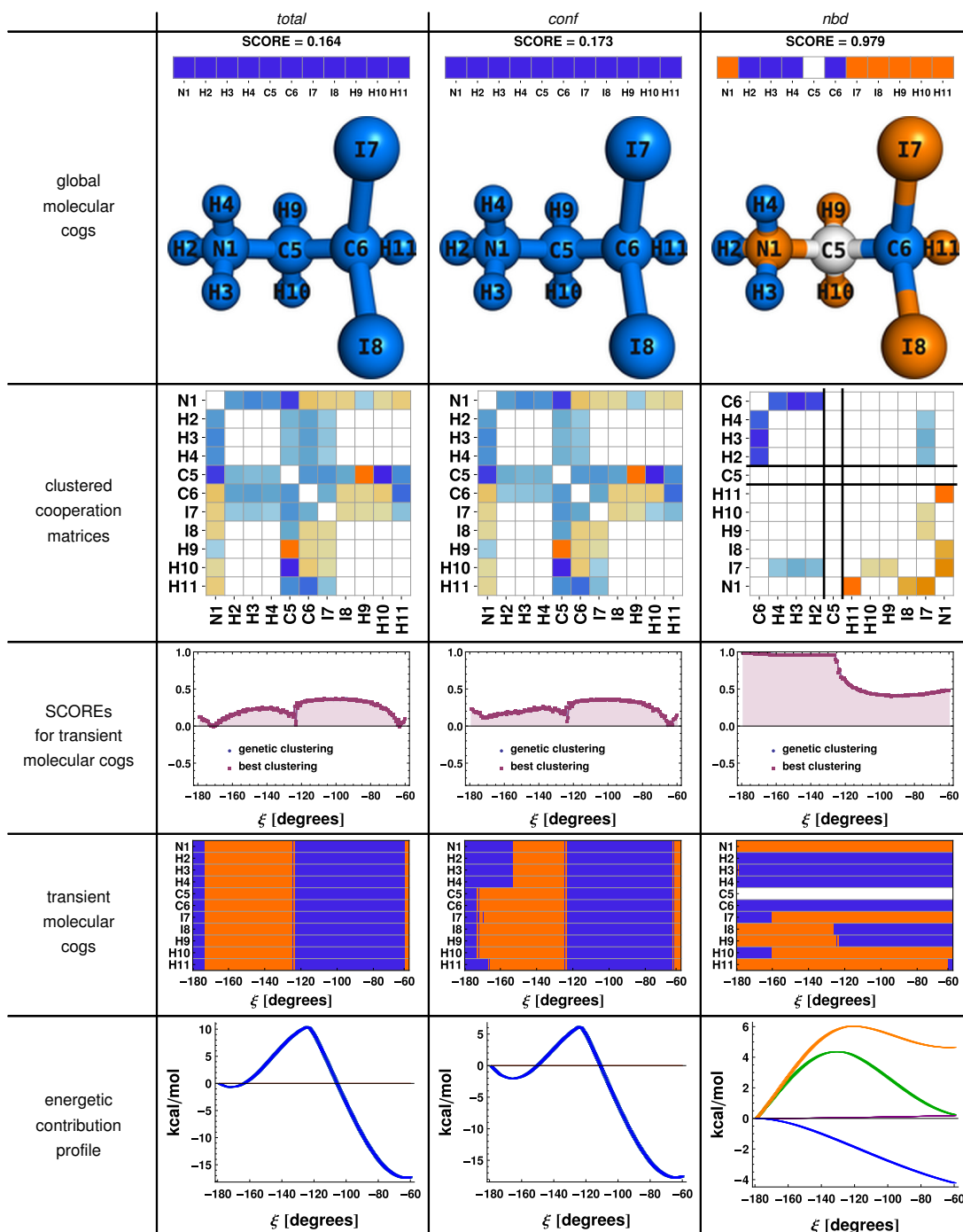


Table 1.2: Results summary for interactions: *total*, *conf* and *nb*. Out of the triple: *total*, *conf* and *nb*, only the last one leads to a high SCORE partitioning, yielding non-trivial molecular cogs. Colors: green, blue, orange and purple in the last row of the table denote contributions from the: whole system, forward cogs, reverse cogs and gear grinding, respectively (see the *Overview* section for details).

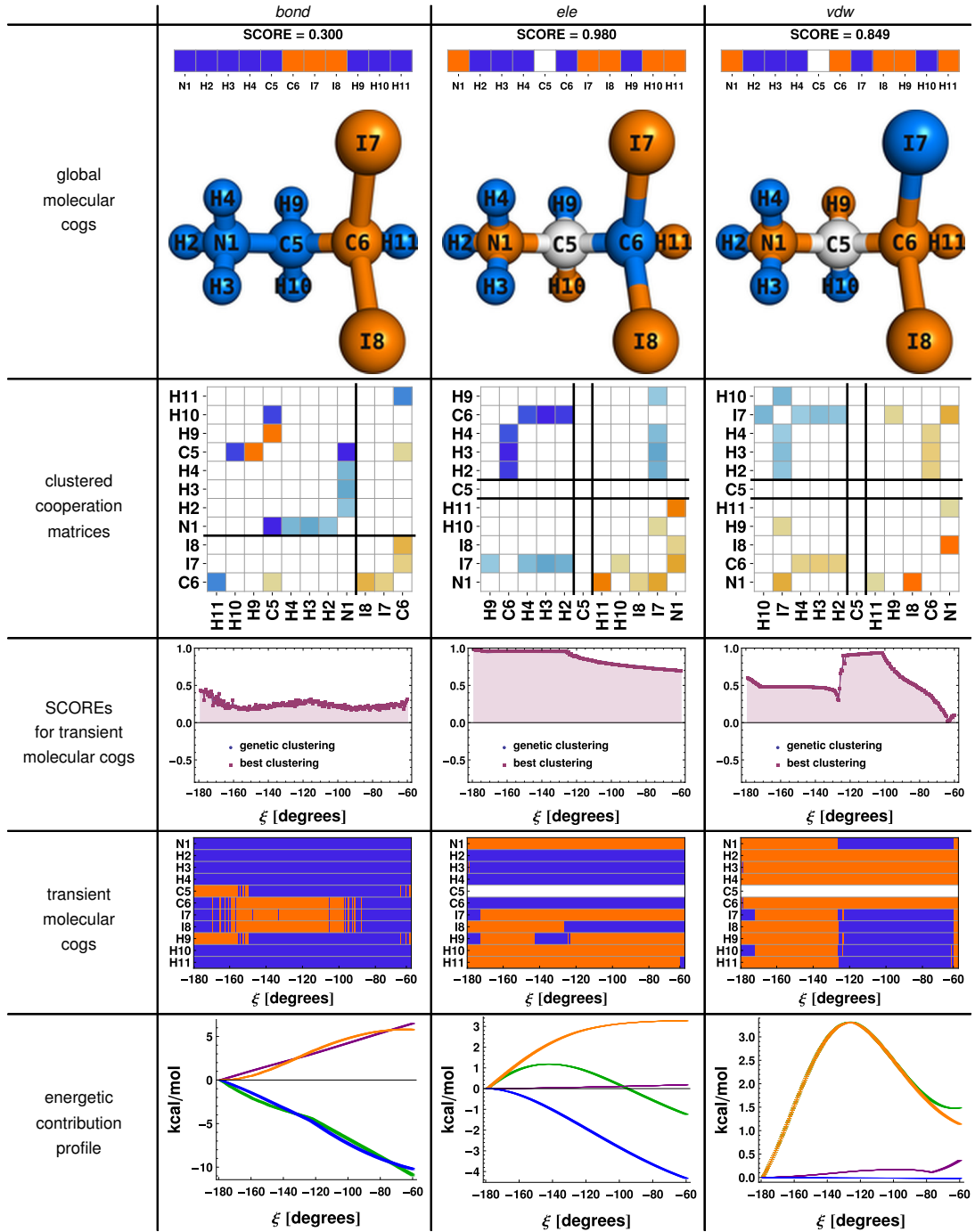


Table 1.3: **Results summary for interactions: *bond*, *ele* and *vdw*.** All three 2-atom interaction types lead to decompositions into non-trivial molecular cogs. However, the transient molecular cogs have high SCOREs only for the *ele* interactions. Colors: green, blue, orange and purple in the last row of the table denote contributions from the: whole system, forward cogs, reverse cogs and gear grinding, respectively (see the *Overview* section for details).

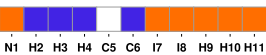
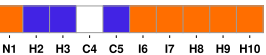
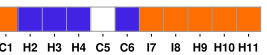
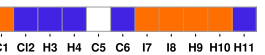
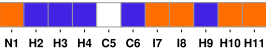
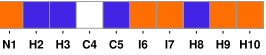
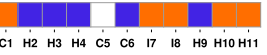
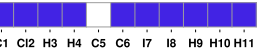
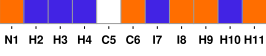
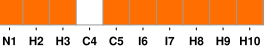
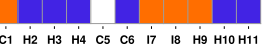
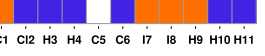
	[NH3+] CC(I)I	NCC(I)I	CCC(I)I	CClCC(I)I
	SCORE = 0.979	SCORE = 0.980	SCORE = 0.995	SCORE = 0.748
<i>nb</i>				
<i>ele</i>	SCORE = 0.980	SCORE = 0.981	SCORE = 0.979	SCORE = 0.381
				
<i>vdw</i>	SCORE = 0.849	SCORE = 0.493	SCORE = 0.989	SCORE = 0.968
				

Table 1.4: **Global molecular cogs for the following molecules:** [NH3+]CC(I)I, NCC(I)I, CCC(I)I and CClCC(I)I. We focused our discussion on partitionings into forward and reverse cogs for the *nb*, *ele* and *vdw* interactions. Full results are detailed in the *Supporting Information*.

Bibliography

- [Aal+97] DM van Aalten et al. “Engineering protein mechanics: inhibition of concerted motions of the cellular retinol binding protein by site-directed mutagenesis.” In: *Protein engineering* 10.1 (1997), pp. 31–37.
- [AAT11] Yelena A. Arnautova, Ruben A. Abagyan, and Maxim Totrov. “Development of a new physics-based internal coordinate mechanics force field and its application to protein loop modeling”. In: *Proteins: Structure, Function, and Bioinformatics* 79.2 (2011), pp. 477–498.
- [Bao09] G Bao. “Protein mechanics: a new frontier in biomechanics”. In: *Experimental mechanics* 49.1 (2009), pp. 153–164.
- [BK95] Stefan Boresch and Martin Karplus. “The meaning of component analysis: decomposition of the free energy in terms of specific interactions”. In: *Journal of molecular biology* 254.5 (1995), pp. 801–807.
- [BS95] G Patrick Brady and Kim A Sharp. “Decomposition of interaction free energies in proteins and other complex systems”. In: *Journal of molecular biology* 254.1 (1995), pp. 77–85.
- [Car+89] EA Carter et al. “Constrained reaction coordinate dynamics for the simulation of rare events”. In: *Chemical Physics Letters* 156.5 (1989), pp. 472–477.
- [Chi+13] Harry Chiang et al. “Molecular mechanics and dynamics characterization of an in silico mutated protein: A stand-alone lab module or support activity for in vivo and in vitro analyses of targeted proteins”. In: *Biochemistry and Molecular Biology Education* 41.6 (2013), pp. 402–408.
- [Chi14] Christophe Chipot. “Frontiers in free-energy calculations of biological systems”. In: *Wiley Interdisciplinary Reviews: Computational Molecular Science* 4.1 (2014), pp. 71–89.
- [Col98] Rowena Marie Cole. *Clustering with genetic algorithms*. Citeseer, 1998.
- [Com+14] Jeffrey Comer et al. “The Adaptive Biasing Force Method: Everything You Always Wanted To Know but Were Afraid To Ask”. In: *The Journal of Physical Chemistry B* (2014).

- [CP07] Christophe Chipot and Andrew Pohorille. *Free energy calculations: theory and applications in chemistry and biology*. Vol. 86. Springer, 2007.
- [Dil97] Ken A Dill. “Additivity principles in biochemistry”. In: *Journal of Biological Chemistry* 272.2 (1997), pp. 701–704.
- [FD07] Brendan J Frey and Delbert Dueck. “Clustering by passing messages between data points”. In: *science* 315.5814 (2007), pp. 972–976.
- [GRC12] James C Gumbart, Benoît Roux, and Christophe Chipot. “Standard binding free energies from computer simulations: What is the best strategy?” In: *Journal of chemical theory and computation* 9.1 (2012), pp. 794–802.
- [Kir35] John G Kirkwood. “Statistical mechanics of fluid mixtures”. In: *The Journal of Chemical Physics* 3.5 (1935), pp. 300–313.
- [LS07] Richard Lavery and Sophie Sacquin-Mora. “Protein mechanics: a route from structure to function”. In: *Journal of biosciences* 32.1 (2007), pp. 891–898.
- [OBo+11] Noel M O’Boyle et al. “Open Babel: An open chemical toolbox”. In: *J Cheminf* 3 (2011), p. 33.
- [Per+12] Xavier Periole et al. “Structural determinants of the supramolecular organization of G protein-coupled receptors in bilayers”. In: *Journal of the American Chemical Society* 134.26 (2012), pp. 10959–10965.
- [Pre07] William H Press. *Numerical recipes 3rd edition: The art of scientific computing*. Cambridge university press, 2007.
- [Sac14] Sophie Sacquin-Mora. “Motions and mechanics: investigating conformational transitions in multi-domain proteins with coarse-grain simulations”. In: *Molecular Simulation* 40.1-3 (2014), pp. 229–236.
- [SG13] Christian Seifert and Frauke Gräter. “Protein mechanics: How force regulates molecular function”. In: *Biochimica et Biophysica Acta (BBA)-General Subjects* 1830.10 (2013), pp. 4762–4768.
- [TV77] Glenn M Torrie and John P Valleau. “Nonphysical sampling distributions in Monte Carlo free-energy estimation: Umbrella sampling”. In: *Journal of Computational Physics* 23.2 (1977), pp. 187–199.
- [WR05] Hyung-June Woo and Benoît Roux. “Calculation of absolute protein–ligand binding free energy from computer simulations”. In: *Proceedings of the national academy of sciences of the united states of america* 102.19 (2005), pp. 6825–6830.