Supporting Information

Strategy To Discover Diverse Optimal Molecules in the Small Molecule Universe

Chetan Rupakheti[1], Aaron Virshup[2], Weitao Yang[2,3], David N. Beratan[2,3,4]

[1]Program in Computational Biology and Bioinformatics, Duke University, Durham NC 27708

[2]Department of Chemistry, Duke University, Durham NC 27708

[3]Department of Physics, Duke University, Durham NC 27708

[4]Department of Biochemistry, Duke University, Durham NC 27708

Corresponding Author

*Phone: (919) 660-1526. E-mail: david.beratan@duke.edu

*Phone: (919) 660-1562. E-mail: weitao.yang@duke.edu

- **SI Definitions**

The *fitness landscape* maps the relation between the Cartesian position of molecular structures and their chemical properties. The landscape can be smooth with a single optimum, or rugged with multiple optima, or flat with very few optima.[1]

The *ruggedness of a fitness landscape* is defined as change in fitness value with respect to change in chemical space distance.[1] Autocorrelation function has been used to quantify ruggedness as a correlation between chemical space distance and fitness value.[1]

The *autocorrelation function* quantifies the correlation of fitness values among binary strings within a hamming distance. It is defined as:[1]

$$\rho(d) \equiv \frac{1}{f_{\mathrm{var}iance}} \frac{1}{N} \sum_{dist(g,g')\in d} (f(g) - f_{mean})(f(g') - f_{mean}) \quad (1)$$

where,

d is the hamming distance between any two binary string g and g'

$\rho(d)$ is the autocorrelation among binary strings within d distance

$f_{\mathrm{var}iance}$ is the variance of the fitness function

f_{mean} is the mean of the fitness function

$f(g), f(g')$ are the fitness values of binary strings g and g'

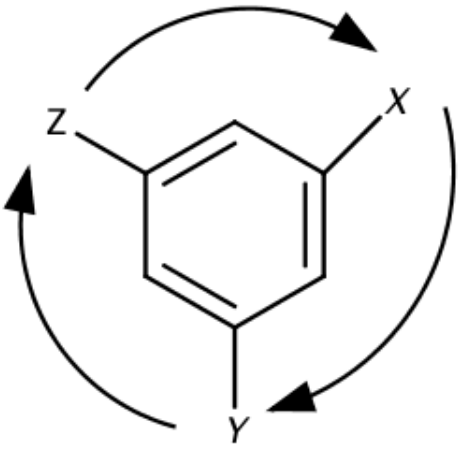
N is the number of binary strings with a hamming distance d

- **SI Method and Results:**

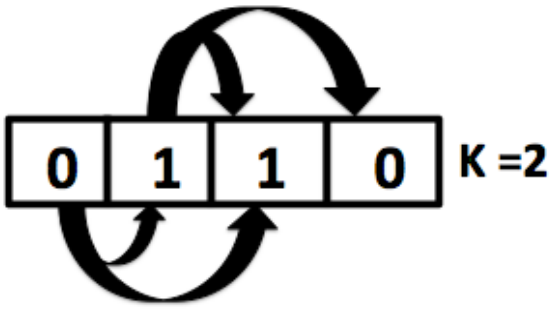
1) Similarity between optimum binary string design using NKp fitness function and small molecule design using ensemble-averaged dipole moment

We found interesting qualitative similarities between small molecule design for target property and design of optimal binary strings using NKp model. The NKp model factors in the influence of each of the *N* different cell, and its possible *k*-wise coupling to other cells (Figures S1A and S1B), on the fitness of a binary string. Such coupling of multiple-sites is seen in actual molecule design problems. Let us image a small molecule design problem with three sites (X, Y, and Z as shown in Figure S1A). A functional

group in a site, for example X, can influence the contribution of functional groups in other sites (Y and Z) to the property value and vice-versa. By *K*-wise linking cells in a binary string, we are modeling the coupling among multiple design positions on the property value. It is also likely that not all functional group make equal contributions towards a molecular property. We can model such uneven fitness contributions by (1) assigning a fitness value of each cell of a binary string randomly from 0 to 1, and (2) further tuning the fitness contribution by using a parameter *p* that sets the fitness contribution to zero for some cells (higher values of p assigns fitness contributions of a larger number of cell to 0). The coupling among the design sites and uneven fitness contributions controlled by parameters *K* and *p* in this model produce a rugged fitness landscape with multiple optima. In fact, previous studies on NKp fitness landscapes indicate that higher values of $K > 0$ increase the ruggedness of the fitness landscape.[1,2] Previous studies also indicate that larger values of *p* flatten the peaks observed in the fitness landscape, producing many structures with a low fitness value (fitness value close to zero).[1,2] Neutral and rugged fitness landscapes have been found in chemical spaces.[3,4]



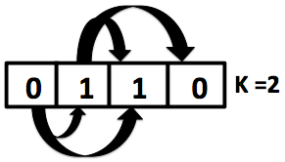
A



B

Figure S1. The similarity between the molecular design problem and the NKp fitness model is indicated. (A) Molecular design where the fitness depends on functional groups X, Y, and Z and their interactions. (B) Binary string model where the fitness contribution from one cell is influenced by two other cells.

A schematic description of the fitness evaluation for the N=4 and K=2 example is shown below.



$$\Phi(0,1,1,0) = \frac{1}{4}\left[\phi_0(0,1,1) + \phi_1(1,1,0) + \phi_2(1,0,0) + \phi_3(0,0,1)\right]$$

A **B**

Figure S2. (A) Example of a binary string of length N=4, where the fitness of a cell is influenced by the composition of its two adjacent cells. (B) Fitness (Φ) of a binary string is the sum of the fitness values for each cell.

The fitness (Φ) of a bit string (g) in the NKp fitness landscape model is:[1]

$$\Phi(g) = \frac{1}{N}\sum_{i=1}^{N}\phi_i(g) \qquad (2)$$

where

$g \in Q^N$ and $Q = 0$ or 1; N = *length of* g

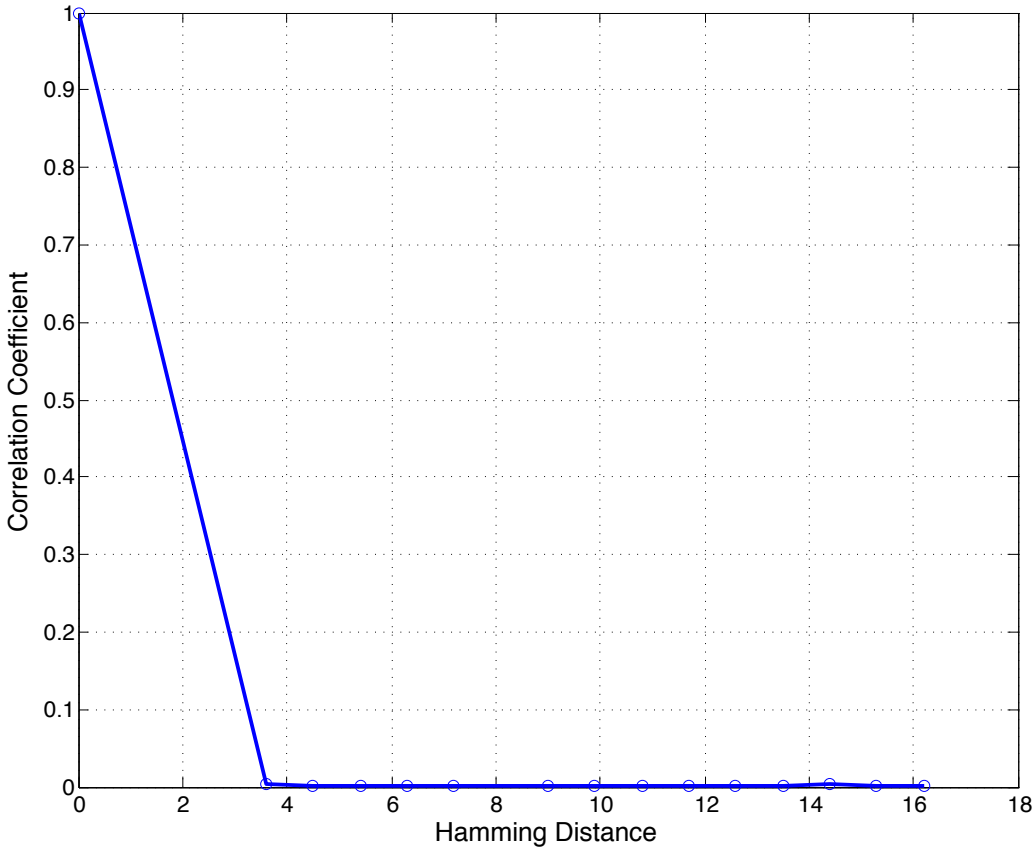
$\phi_i \in [0,1]$, where ϕ_i is drawn randomly from [0,1]

Mimicking GDB-9 dipole moment landscape with NKp landscape model:

We vary parameters K (0 to N-1) and p (0 to 1) to construct a fitness landscape similar to the GDB-9 landscape. For our study, we fixed N that leads to the size of search space comparable to GDB-9 i.e. 524288 bit strings. We tried p from the range 0 to 1 to mimic the distribution of GDB-9 dipole moment property (figures 2A and 2B). The

parameter *K* from 0 to *N*-1 was chosen to mimic the autocorrelation function (eq. 1) plotted on figures S3 A and B. The autocorrelation function (eq. 1) was used to quantify the qualitative similarity between the two landscapes. We found that the fitness landscape corresponding to *N=19, K=9,* and *p=0.9* has similar autocorrelation characteristics to GDB-9 (Figures S3 A and B).

A



B

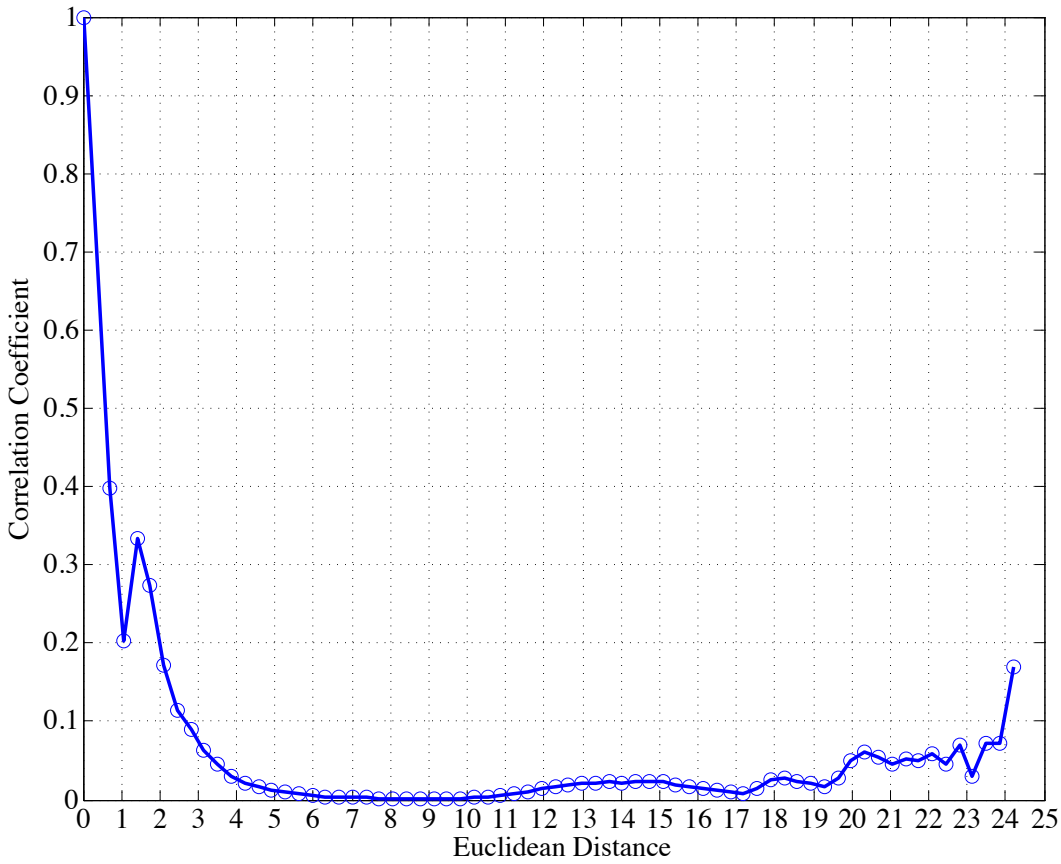


Figure S3. Autocorrelation plots for the NKp (*N*=19, *K*=9, *p*=0.9) model landscape (A) and GDB-9 space (B). The correlation between distance and fitness (the autocorrelation decays to zero in 4 distance units) in the NKp and molecular dipole spaces.

The correlation between property and molecular distance decays rapidly in both landscapes (see Figures S3 A and B). However, the distance metric in the NKp model problem is the Hamming distance (arising from the use of a binary strin), while the distance metric in the molecular library is the Euclidean (computed using the molecular autocorrelation descriptors). Figure S3B indicate that the correlation between the molecular dipole moment and the autocorrelation descriptor distances does not decay linearly with distance. This behavior corresponds to the NKp case in Figure S3A where

K>0. We see from these comparisons that the GDB-9 molecular property landscape is rugged for the molecular descriptors used here.

2) How does the landscape change with ruggedness parameter '*K*'?

It is known that the ruggedness of a fitness landscape increases with the increase in parameter '*K*'.[1] To demonstrate this we chose N=10 and varied K =0 to N-1. Autocorrelation function was applied to compute the correlation between hamming distance and NKp fitness value. The result is shown below:

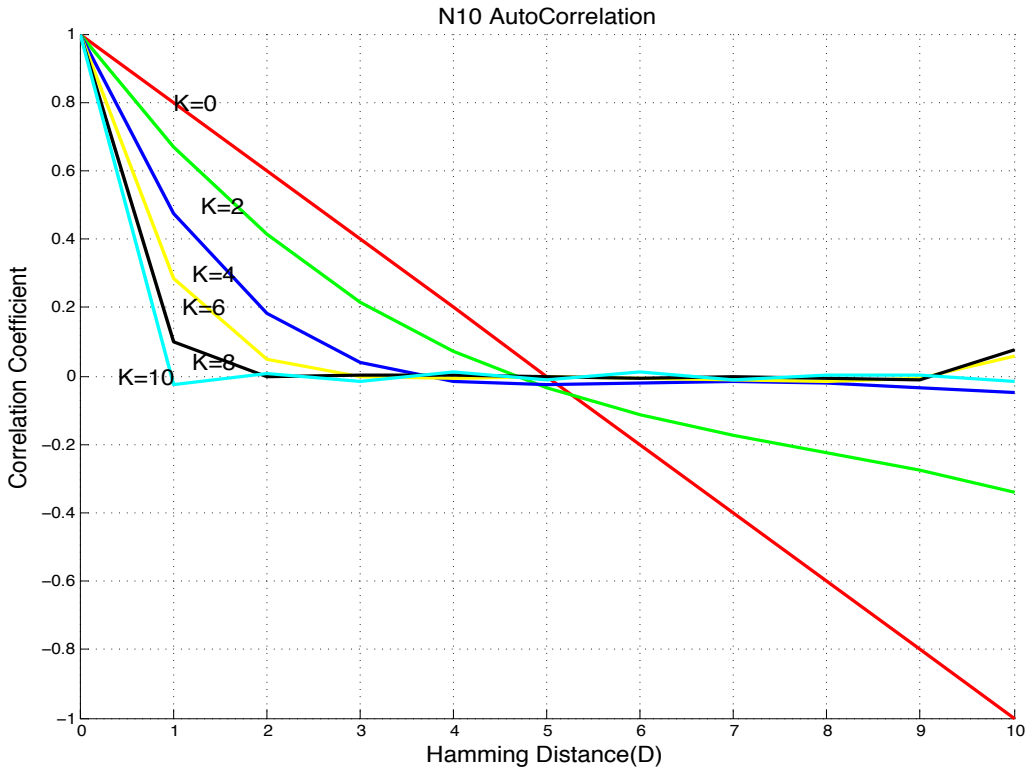


Figure S4. Compares plot of autocorrelation function with varying hamming distance for *N*=10 where *K*=0 to *N*

From Figure S4 we notice that when K=0 (no coupling among sites), we observe linearly decaying autocorrelation function with respect to hamming distance. This behavior is representative of smooth fitness landscape where the fitness value decays smoothly with respect to chemical space distance.[1] As for K >0 cases, the autocorrelation function starts to decay rapidly with respect to hamming distance.[1] This behavior is representative of rugged fitness landscape where a small change in chemical space distance results in large change in fitness value (also known as activity cliff).[4]

3) Run times involved in property-optimizing ACSESS calculation

Steps Involved	**CPU Time**
Conformer Generation	5 Sec/conformer
Descriptor Calculation	4 Sec/molecule
Semi-Empirical AM1 Calculation involving: • Geometry optimization • Dipole moment calculation	30 Sec/conformer
ACSESS algorithm Involving following steps: • Mutation • Crossover • Diversity Calculation • GDB Filtering	330 Sec/100 iterations of ACSESS

- **Reference**

(1) Barnett, L. Ruggedness and Neutrality - The NKp Family of Fitness Landscapes. **1997**.

(2) Geard, N.; Wiles, J.; Hallinan, J.; Tonkes, B.; Skellett, B. A Comparison of Neutral Landscapes - NK, NKp and NKq. *Proc. 2002 Congr. Evol. Comput. CEC'02 (Cat. No.02TH8600) 1*, 205–210.

(3) Wassermann, A. M.; Wawer, M.; Bajorath, J. Activity Landscape Representations for Structure-Activity Relationship Analysis. *J. Med. Chem.* **2010**, *53*, 8209–8223.

(4) Bajorath, J.; Peltason, L.; Wawer, M.; Guha, R.; Lajiness, M. S.; Van Drie, J. H. Navigating Structure-Activity Landscapes. *Drug Discov. Today* **2009**, *14*, 698–705.