SUPPORTING INFORMATION

Supplementary figures and tables

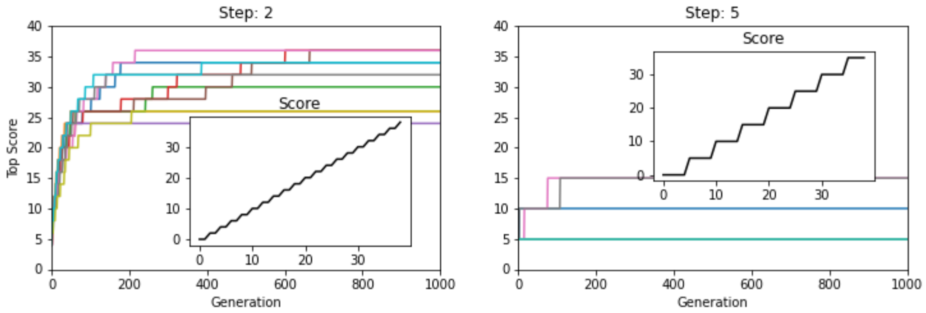

Figure S1. Similar to the Shakespeare example but with a discontinuous score (inset): score = (correct//step)*step, where // denotes Python integer division, correct is the number of correct characters in the sequence and step = 2 and 5, respectively. The insets show a plot of score vs correct.

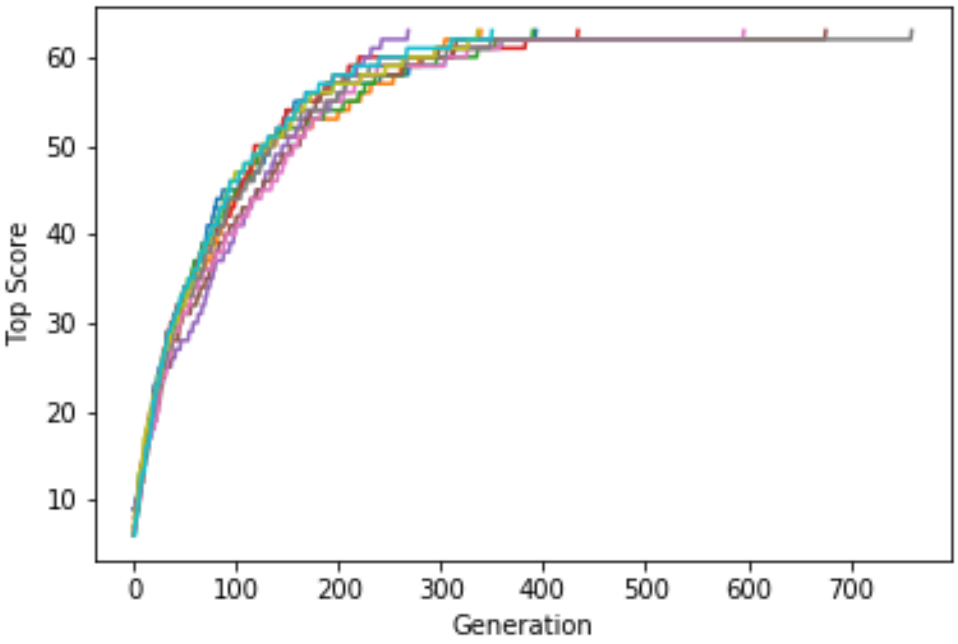

Figure S2. Similar to the Shakespeare example but with the phrase ”to be or not to be that is the question whether it is nobler in”, which contains the same number of characters (61) as the SMILES string for tiotixene.

Table S1. Non-canonical SMILES strings from successful SMILES-based GA searchers. In the case of troglitazone none of the searchers were successful, so SMILES with a Tanimoto similarity of $0.\overline{79}$ are shown. Notice that most of the SMILES strings, which have not been canonicalized, for each molecule have the general syntax.

Celecoxib
NS(=O)(=O)C1=CC=C(N2-N=C(C(F)(F)F)C=C2C/2=CC=C(C)C=C2)C=C1
NS(=O)(=O)C1=CC=C(N2N=C(C(F)(F)F)C=C(C3=CC=C(C)C=C3)2)[C@H]=C1
NS(=O)(=O)C1=[C@H]C=C(N2N=C(C(F)(F)F)C=C2C2=CC=C(C)C=C2)C=C1
NS(=O)(=O)C1=CC=C(N2N=C(C(F)(F)F)C=C2C2=CC=C(C)C=C2)[C@@H]=C1
NS(=O)(=O)C1=CC=C(N2N=C(C(F)(F)F)C=C2C2=CC=C(C)C=C2)C=C1
CC1=CC=C(N2N=C(C(F)(F)F)C=C2C2=CC=C(S(N)(=O)=O)C=C2)C=C1
NS(=O)(=O)C1=CC=C(N2N=C(C(F)(F)F)C=C2C2=[C@@H]C=C(C)C=C2)C=C1
NS(=O)(=O)C1=CC=C(N2N=C(C(F)(F)F)C=C2C2=CC=C(C)C=C2)C=C1
N1=C(C(F)(F)F)[C@H]=C(C2=CC=C(S(N)(=O)=O)C=C2)N1C1=CC=C(C)C=C1
NS(=O)(=O)C1=CC=C(N2N=C(C(F)(F)F)C=C2C2=CC=C(C)C=C2)[C@H]=C1
NS(=O)(=O)C1=CC=C(N2N=C(C(F)(F)F)C=C2C2=CC=C(C)C=C2)C=C1
NS(=O)(=O)\\C1=C[C@@H]=C(N2N=C(C(F)(F)F)C=C2C2=CC=C(C)C=C2)C=C1
NS(=O)(=O)C1=CC=C(N2N=C(C(F)(F)F)C=C(C3=CC=C(C)C=C3)2)C=C1
NS(=O)(=O)C1=[C@H]C=C(N2N=C(C(F)(F)F)C=C2C2=CC=C(C)C=C2)C=C1
NS(=O)(=O)C1=CC=C(N2N=C(C(F)(F)F)C=C(C3=CC=C(C)C=C3)2)C=C1
NS(=O)(=O)C1=CC=C(N2N=C(C(F)(F)F)C=C2C2=CC=C(C)C=C2)C=[C@@H]1
NS(=O)(=O)C1=CC=C(N2N=C(C(F)(F)F)C=C2\C2=CC=C(C)C=C2)C=C1
NS(=O)(=O)C1=CC=C(N2N=C(C(F)(F)F)C=C2C2=CC=C(C)[CH]=C2)C=C1
Troglitazone
CC1=C(C)C2=C(CCC(C)(C[O@]C)O2)C(C-OC2=CC=C(C[C@H]3SC(=O)NC3=O)C=C2)=C1O
C1=C(-OCC2=C(O)C(C)=C(C)C3=C2CCC(COC)(C)O3)C=CC(C[C@@H]2SC(=O)NC2=O)=C1
CC1=C(O)C(C)=C(COC2=CC=C(C[C@@H]3SC(=O)NC3=O)C=C2)C2=C1CCC(COC)(C)O2
OC1=C(COC2=CC=C(C[C@H]3SC(=O)NC3=O)C=C2)C(C)=C2OC(C)(COC)CCC2=C1C
CC1=C(COC2=CC=C(CC3SC(=O)NC3=O)C=C2)C2=C(CCC(COC)(C)O2)C(C)=C1O
OC1=C(C)C(C)=C2OC(COC)(C)C-CC2=C(COC2=CC=C(C[C@H]3SC(=O)NC3=O)C=C2)\1
CC1=C(C)C2=C(/CCC(C[O]C)(C)O2)C(COC2=CC=C(CC3SC(=O)NC3=O)C=C2)=C1O
CC1=C(C\OC2=CC=C(C[C@H]3SC(=O)NC3=O)C=C2)C2=C(CCC(C)(COC)O2)C(C)=C1O
CC1=C(COC2=CC=C(CC3SC(=O)NC3=O)C=C2)C2=C(CCC(C)(COC)O2)C(C)=C1O
CC1=C(COC2=CC=C(C[C@H]3SC(=O)NC3=O)C=C2)C2=C(CCC(C)(COC)O2)C(C)=C1O
[C@@H]1=C(C[C@H]2SC(=O)NC2=O)C=[C@H]C(OCC\2=C(O)C(C)=C(C)C3=C2CCC(C)(\COC)O3)=[C@H]
CC1=C(C)C2=C(CCC(C)(CO\O)O2)C(C)=C1COC1=CC=C(C[C@@H]2SC(=O)NC2=O)C=C1
CC1=C(C)C2=C(CCC(C)(COO)O2)C(C)=C(COC2=CC=C(CC3SC(=O)NC3=O)C=C2)1
OC1=C(-COC2=CC=C(CC3SC(=O)NC3=O)C=C2)C(C)=C2OC(COC)(C)CCC2=C1C
CC1=C(C)C2=C(CCC(C)(COC)O2)C(COC2=CC=C(C[C@H]3SC(=O)NC3=O)C=C2)=C1O
Tiotixene
CN(C)S(=O)(=O)C1=CC=C2SC3=CC=CC=C3C(=CC\C(N3CCN(C)CC3))C2=C1
CN(C)S(=O)(=O)C1=CC=C2SC3=CC=[C@H]C=C3C(=C(CCN3CCN(C)CC3))C2=C1
CN(C)S(=O)(=O)C1=CC=C2SC3=CC=CC=C3C(=[C@H]C\C(N3CCN(C)CC3))C2=C1

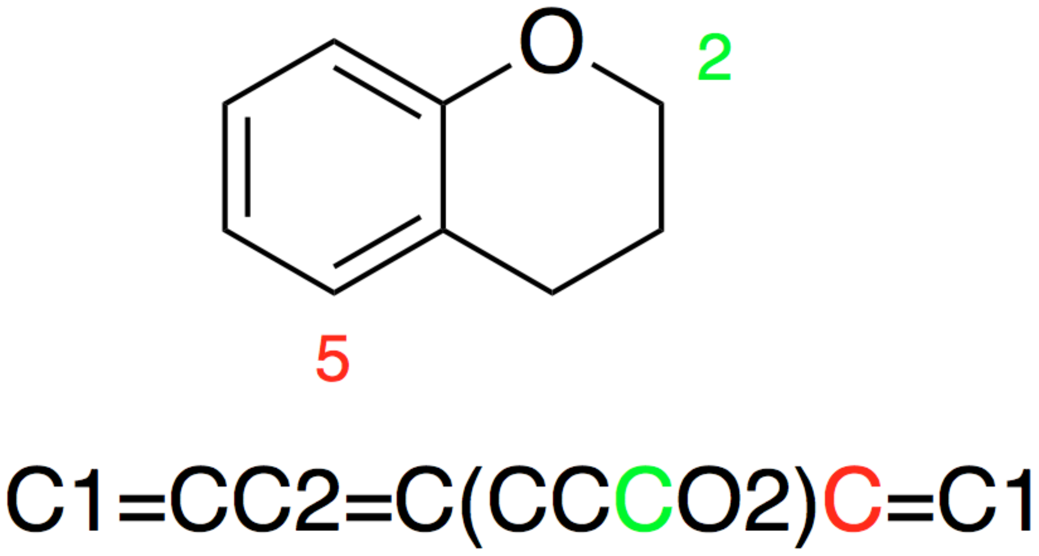

Figure S3. The SMILES syntax for the chromane moiety that is most commonly found in the initial population. Since the 2 position is sandwiched between 2 ring closures (denoted by ’1’ and ’2’) the benzylthiazolidinedione group must have the form COC3=CC=C(CC4SC(=O)NC4=O)C=C3, while it can have the form COC2=CC=C(CC3SC(=O)NC3=O)C=C2 at the 2 position. However, only 27% of the SMILES strings in the initial population contain with ’4’, compared with 72% for ’3’.

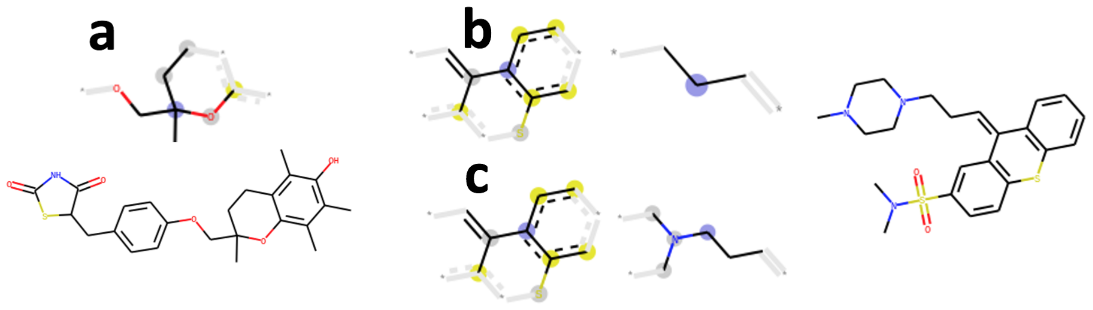

Figure S4. (a) This fragment is not present in the initial population used for rediscovery of troglitazone (shown below the fragment). (b) and (c) These two fragments are not present in the initial population used for rediscovery of titoxene (shown to the right of the fragments). For (a) and (b) the initial populations are constructed by screening 1.6 million molecules, while for (c) the initial populations are constructed by screening 10,000 molecules.

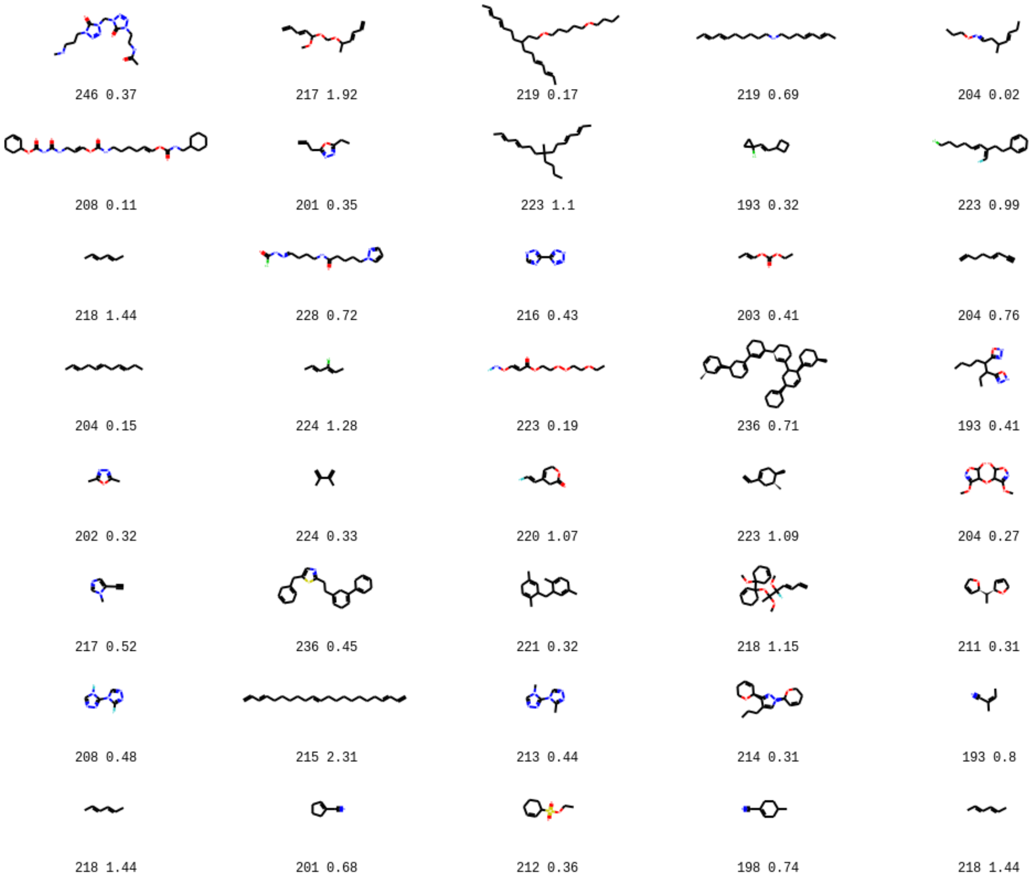

Figure S5. Molecules that absorb near 200 nm.

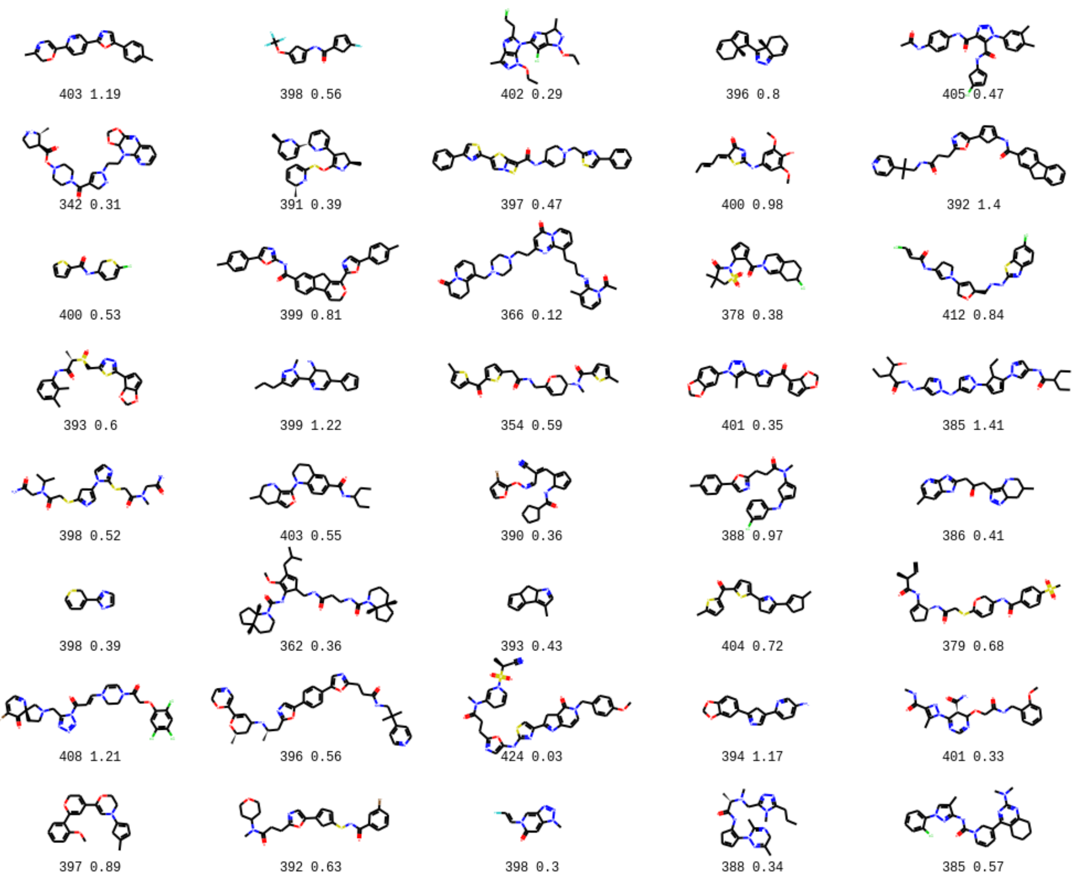

Figure S6. Molecules that absorb near 400 nm. The values are recalculated so any deviation >7 nm from the target wavelength is due to conformational effects.

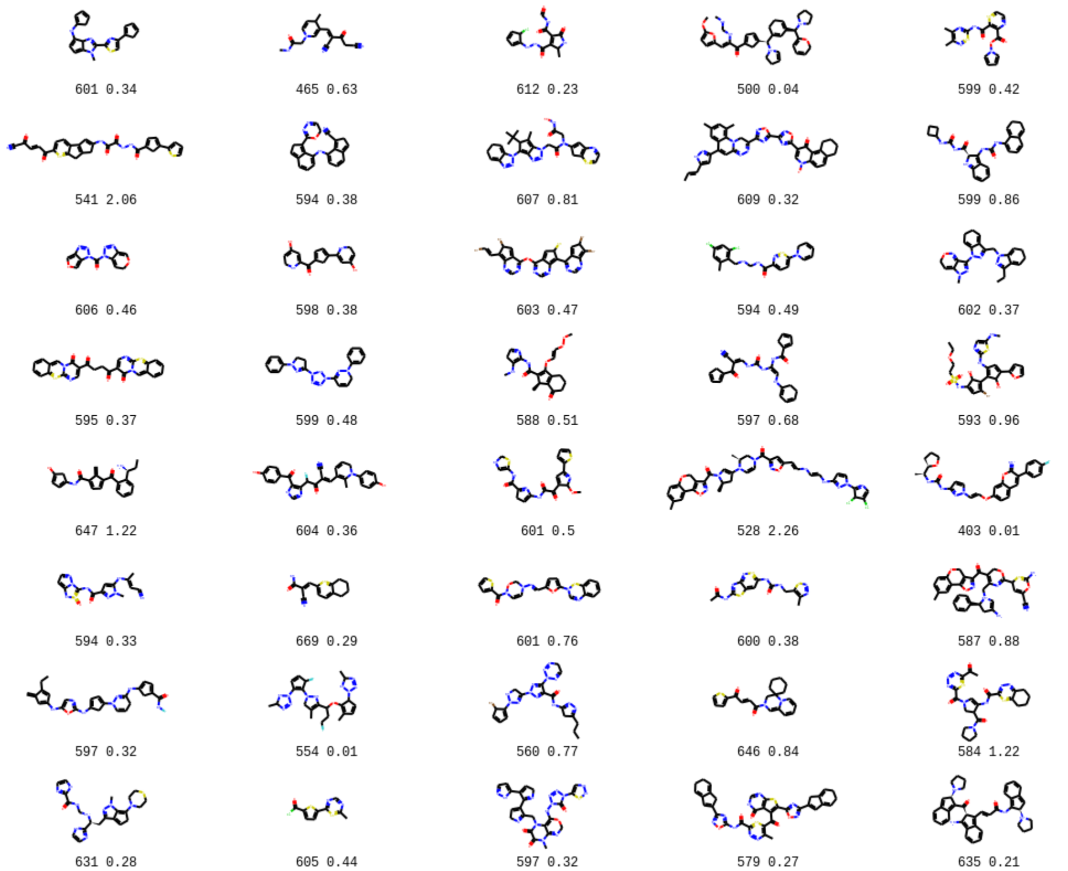

Figure S7. Molecules that absorb near 600 nm. The values are recalculated so any deviation >7 nm from the target wavelength is due to conformational effects (with the exception of one case).

Number of sequences with one correctly place letter

If there are c different characters and l positions, then the number of possible character sequences of length l is c^l. There are $c-1$ incorrect characters for each of the l positions, or $(c-1)^l$ possible sequences with all characters placed incorrectly. Thus, there are $c^l-(c-1)^l$ sequences with at least one character placed correctly, which corresponds to a probability of

$$p = 1 - \left(\frac{c-1}{c}\right)^l \tag{S1}$$

For c = 27 and l = 39, p = 0.77

Number of correctly placed characters present in the initial population

If N is the number of randomly chosen phrases of length l in the initial population and each position in the phrase has a probability $p = 1/c$ of being correct in a random phrase then the probability that a character is placed correctly in n out of N random sequences is given by:

$$P(n;N,p) = \binom{N}{n} p^n (1-p)^{(N-n)} \tag{S2}$$

$P(n=0;N,p=1/27)$ is the probability that a character is not placed correctly in any of the N sequences. The probability, that a character at a certain position is placed correctly in at least one of the 100 phrases is then

$$p_1 = 1 - P(n=0;N=100,p=1/27) = 0.977 \tag{S3}$$

The probability that x of the 39 characters are placed correct somewhere in the 100 phrases is then described by a new Binomial distribution with number of tries, l=39, and probability of success, p_1=0.977. Using properties of the Binomial distribution, the average number of correctly place characters in the initial population.

$$\langle x \rangle = l p_1 = 38.1 \tag{S4}$$

and the standard deviation:

$$\sigma_x = \sqrt{l p_1 (1-p_1)} = 0.9 \tag{S5}$$

Number of GA runs needed to achieve >99% certainty of success

If the probability of a successful search is p, then the probability of N searchers all failing is $(1-p)^N$, and the probability that at least one search succeeds is $1-(1-p)^N$ Thus, if we want a greater than 99% certainty of at least one successful search $(0.99 > 1-(1-p)^N))$, then the minimum number of searches that must be performed is

$$N = \left\lceil \frac{\ln(0.01)}{\ln(1-p)} \right\rceil \tag{S6}$$

where $\lceil x \rceil$ indicates that x is rounded up to the nearest integer.