Electronic Supplementary Information for "Generative Organic Electronic Molecular Design Informed by Quantum Chemistry"

Cheng-Han Li and Daniel P. Tabor*

Department of Chemistry, Texas A&M University, College Station, TX 77842 USA

E-mail: daniel_tabor@tamu.edu

Contents

S1 Description of Contents of Raw Data Files

All the code and data, including the modified REINVENT packages, initial ChEMBL prior network, scripts for setting up configurations and running REINVENT, calibration set along with TD-DFT results and references, and generated molecules in this study are provided at the following GitHub repository:

https://github.com/Tabor-Research-Group/reinvent_qc.

S2 Implementation Details for REINVENT

The REINVENT learning cycle mainly consists of two parts, agent and score modulating block.[S1,S2] The agent network samples the SMILES string from building sequences of tokens. The score modulating block evaluates the augmented score for a batch of SMILES sampled by the agent network. The scoring modulating block is made up of prior, scoring function, and diversity filter. The negative log-likelihood (NLL) to sample each SMILES string X from prior would be evaluated as following,

$$\mathrm{NLL}_{\mathrm{Prior}}(X) = -\sum_{i=1}^{N} logP(T_i|T_{i-1}...T_1) \tag{1}$$

where $P(T_i|T_{i-1}...T_1)$ is the probability of sampling the token T_i at step i given the tokens sampled at previous steps. For each valid SMILES sampled by the agent, it would be given a score in the range [0,1] based the the scoring function S. However, a SMILES string would be given a zero if that SMILES string is sampled before or there are too many similar molecules sampled depending on the diversity filter. Finally, the loss function can be calculated as the squared difference between the augmented and agent likelihood.

$$\mathrm{NLL}_{\mathrm{Augmented}}(X) = \mathrm{NLL}_{\mathrm{Prior}}(X) - \sigma \times S(X) \tag{2}$$

$$loss = [\mathrm{NLL}_{\mathrm{Augmented}}(X) - \mathrm{NLL}_{\mathrm{Agent}}(X)]^2 \tag{3}$$

The agent network is updated at each epoch such that the loss function is minimized.

For the first-step curriculum learning, we ran 1,000 iterations with a batch size of 128 molecules, sigma scalar factor of 128, learning rate of 0.001. We implemented diversity filter with a bucket size of 16 and identify molecules with the same BemisMurcko scaffold[S3] in the same bucket if all cheminformatics criteria are satisfied. For the second-step curriculum learning, we ran 500 iterations with a batch size of 64 molecules, sigma scalar factor of 128, learning rate of 0.001. We implemented diversity filter with a bucket size of 16 and identify molecules with the same carbon skeleton dervied from the BemisMurcko scaffold[S3] in the same bucket with a minscore of 0.5. Inception was implemented in this step with a memory size of 100 and sample size of 10.

Here, we provide the list of SMARTS strings and show unrealistic substructures from our initial searches that are used to exclude forbidden substructures:

```
"[*;r3]", "[*;r4]", "[*;r7]", "[*;r8]", "[*;r9]", "[*;r10]", "[*;r11]",
"[*;r12]", "[*;r13]", "[*;r14]", "[*;r15]", "[*;r16]", "[*;r17]", "[#7]~[#7]",
"N=c1[nH]cccc1", "N=c1ccc[nH]1", "c1onccc1", "[#8]~[#8]", "[#7]~[#8]", "[#6;+]",
"[#16][#16]", "[#7;!n][S;!$(S(=O)=O)]", "[#7;!n][#7;!n]",
"[#7;!n][C;!$(C(=[O,N])[N,O])][#16;!s]", "[#7;!n][C;!$(C(=[O,N])[N,O])][#7;!n]",
"[#7;!n][C;!$(C(=[O,N])[N,O])][#8;!o]", "[#8;!o][C;!$(C(=[O,N])[N,O])][#16;!s]",
"[#8;!o][C;!$(C(=[O,N])[N,O])][#8;!o]", "[#16;!s][C;!$(C(=[O,N])[N,O])][#16;!s]",
"C=c1ccccc1=C", "N=c1cc-ccc1", "Cn1cccc1=C", "Cn1ccc(=C)c1", "n1cccc1=C",
"n1ccc(=C)c1", "C1C=c2ccccc2=N1", "c1sc(=N)[nH]c1", "c1[nH]c(=O)ccc1",
"N=c1nccc[nH]1", "N=c1ncc[nH]1", "[*]=N~[*]", "[*;+]", "[*;-]",
"[*]~1~[*]~[*]=[#6]=[*]~[*]~1", "[*]~1~[*]=[#6]=[*]~[*]~1"
```

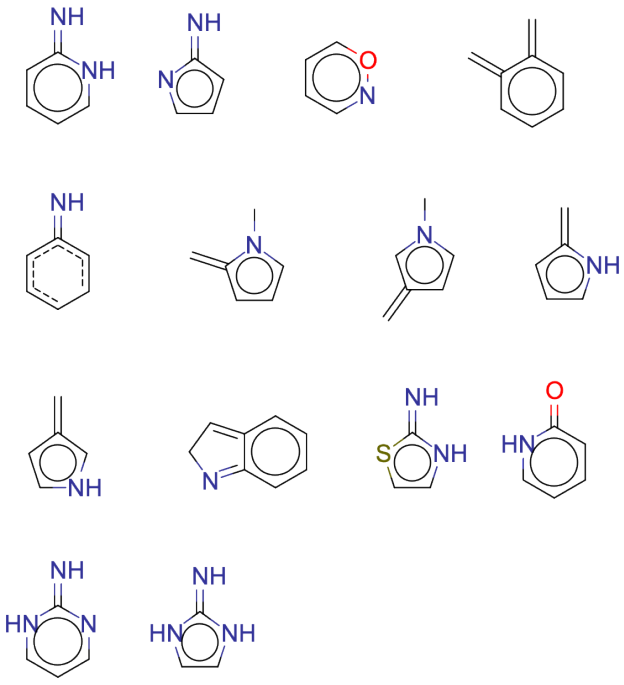

Fig. S1: Forbidden substructures generated from initial REINVENT searches.

S3 Comparison of SCScore between Molecules from REINVENT and Calibration Set

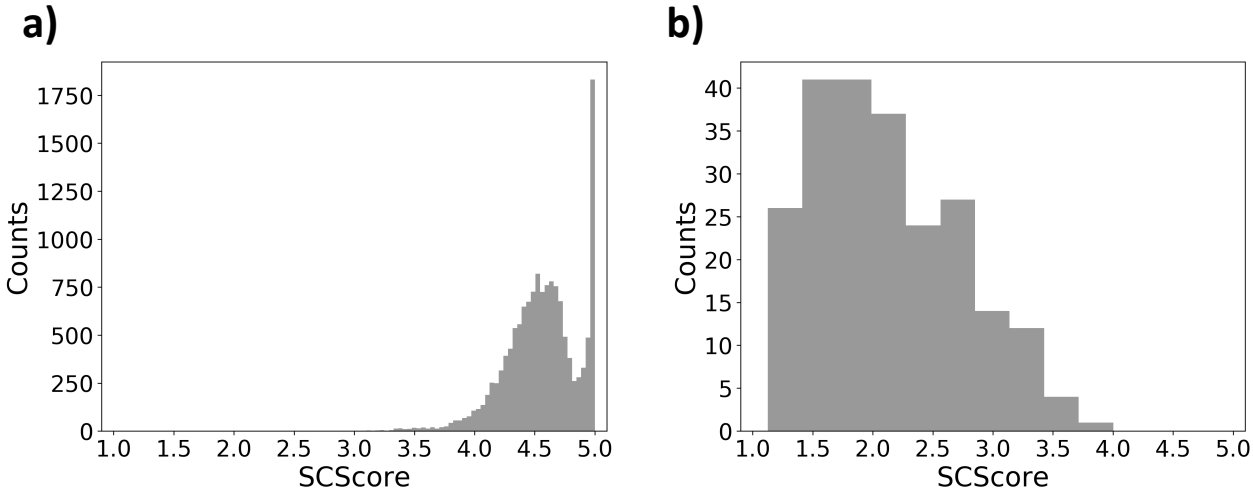

Fig. S2: Distributions of the SCScore for a) top-scoring molecules generated by REINVENT without restrictions on synthesizability and b) molecules in the calibration set for the TD-DFT benchmark study.

S4 Scoring Transformation for Excited-state Energy Gaps

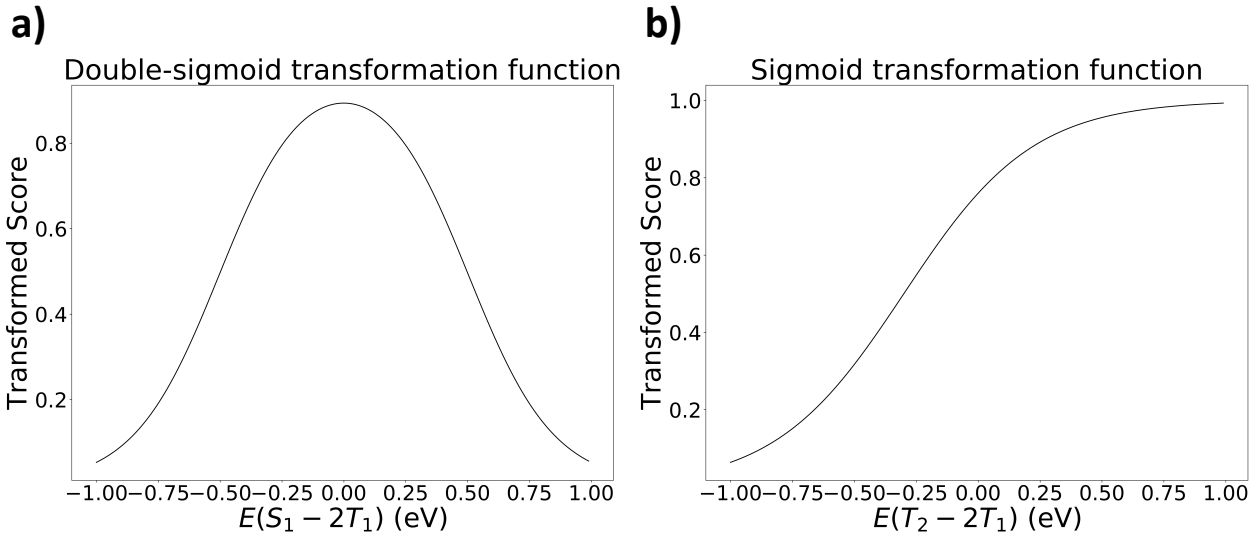

Fig. S3: a) Double-sigmoid transformation for S_1/T_1 energy gap and b) sigmoid transformation for T_2/T_1 energy gap

S5 TD-DFT Benchmark Results

Table S1: Average walltime (second) for excited-state methods

Environment	Ground State Optimization	B3LYP/6-31G(d)	ωB97X-D/def2-SV(P)
Vacuum	GFN2-xTB	41	133
	PBEh-3c	168	242
Implicit Toluene	GFN2-xTB	34	328
	PBEh-3c	102	412

Table S2: Mean absolute error (eV) of S_1/T_1 energies for excited-state methods

Environment	Ground State Optimization	B3LYP/6-31G(d)	ωB97X-D/def2-SV(P)
Vacuum	GFN2-xTB	0.13/0.12	0.15/0.12
	PBEh-3c	0.13/0.12	0.15/0.12
Implicit Toluene	GFN2-xTB	0.13/0.13	0.15/0.12
	PBEh-3c	0.13/0.13	0.15/0.12

S6 Top-10 Average Electronic Scores for Exploratory and Exploitative Cases

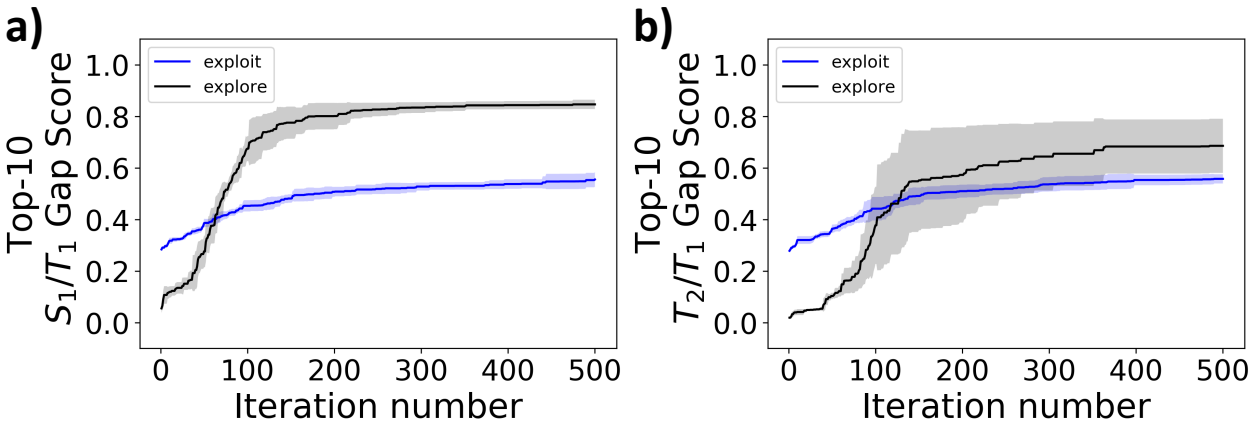

Fig. S4: The optimization curves of top-10 average on a) S_1/T_1 gap score and b) T_2/T_1 gap score.

Table S3: AUC top-10 of S_1/T_1 gap score and T_2/T_1 gap score for exploratory and exploitative cases

	Exploratory Case	Exploitative Case
S_1/T_1 gap score	0.722±0.034	0.490±0.019
T_2/T_1 gap score	0.525±0.122	0.494±0.024

S7 Excited-State Analysis for SF/TTA Candidates

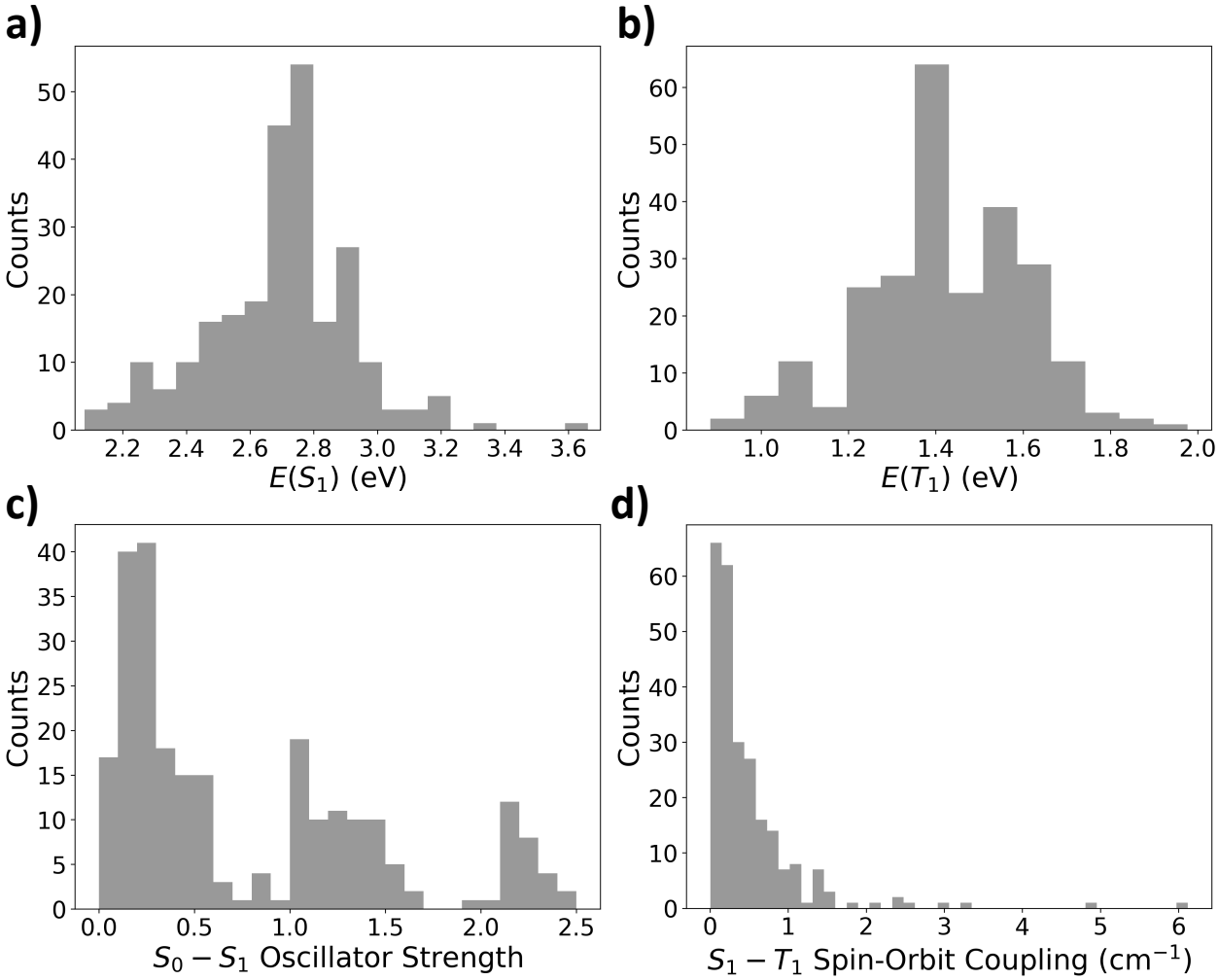

Fig. S5: Distribution of a) S_1 energies, b) T_1 energies, c) $S_0 - S_1$ oscillator strength, and d) $S_1 - T_1$ spin-orbit couplings of SF/TTA candidates meeting both S_1/T_1 and T_2/T_1 criteria.

References

[S1] Olivecrona, M.; Blaschke, T.; Engkvist, O.; Chen, H. Molecular de-novo design through deep reinforcement learning. *J. Cheminform.* **2017**, *9*, 48.

[S2] Blaschke, T.; Arús-Pous, J.; Chen, H.; Margreitter, C.; Tyrchan, C.; Engkvist, O.; Papadopoulos, K.; Patronov, A. REINVENT 2.0: An AI Tool for De Novo Drug Design. *J. Chem. Inform. Model.* **2020**, *60*, 5918–5922, PMID: 33118816.

[S3] Bemis, G. W.; Murcko, M. A. The Properties of Known Drugs. 1. Molecular Frameworks. *J. Med. Chem.* **1996**, *39*, 2887–2893, PMID: 8709122.