Electronic Supplementary Information

From orbital analysis to active learning: an integrated strategy for the accelerated design of TADF emitters

Jean-Pierre Tchapet Njafa,[*a] **Steve Cabrel Teguia Kouam,**[b] **Patrick Mvoto Kongo,**[a] **and Serge Guy Nana Engo**[a]

DOI: 10.1039/b000000x

Contents

S1 Computational Details

S1.1 Software versions and environment

All calculations were performed using the following software packages:

Quantum chemistry tools:

- **xTB** version 6.7.1 – Extended tight-binding calculations (GFN2-xTB)
- **CREST** version 3.0.2 – Conformational search
- **sTDA** version 1.6.3 – Simplified TD-DFT/TDA excited states
- **Multiwfn** version 3.8(dev), 2025-Jul-12 – Wavefunction analysis and NTO generation

Machine learning environment:

- **Python** 3.12.3 – ML pipeline and data processing
- **NumPy** 2.0.2 – Numerical computations
- **pandas** 2.2.3 – Data manipulation
- **scikit-learn** 1.7.2 – Machine learning models (RF, SVR, GB)
- **SHAP** 0.50.0 – Model interpretability
- **Matplotlib** 3.9.3 – Visualization
- **SciPy** 1.14.1 – Scientific computing

[a] *Department of Physics, Faculty of Science, University of Yaounde I, P.O.Box 812, Yaounde, Cameroon. E-mail: jean-pierre.tchapet@facsciences-uy1.cm*
[b] *Department of Physics, Faculty of Science, University of Douala, Po. Box 24157, Douala, Cameroon.*

S1.2 Ground-state geometry optimization

Ground-state geometries were optimized using the GFN2-xTB method in two environments:

Gas phase:

```
xtb molecule.xyz --opt tight
```

Toluene solvent:

```
xtb molecule.xyz --opt tight --gbsa toluene
```

Key parameters:

- Optimization convergence: `tight` ($\Delta E < 10^{-6}$ E$_h$)
- Gas phase: vacuum calculations without implicit solvation
- Implicit solvation: GBSA model (toluene, $\varepsilon = 2.38$)
- SCF convergence: 10^{-8} E$_h$

Both environments were considered to evaluate the solvent effects on TADF properties.

S1.3 Excited-state calculations

Excited states were computed using the sTDA and sTD-DFT-xTB methods as implemented in the `stda` program (version 1.6.3). Both methods read the xTB wavefunction file (`wfn.xtb`) automatically from the working directory.

sTDA-xTB (Simplified Tamm-Dancoff Approximation):

```
stda -xtb -e 10      # singlets
stda -xtb -e 10 -t   # triplets
```

sTD-DFT-xTB (Simplified TD-DFT with RPA):

```
stda -xtb -rpa -e 10      # singlets
stda -xtb -rpa -e 10 -t   # triplets
```

The `-xtb` flag enables the use of GFN-xTB orbitals, while `-rpa` switches from the Tamm-Dancoff approximation (TDA) to the full random phase approximation (RPA), i.e., sTD-DFT. The `-t` flag computes triplet states instead of singlets.

Parameters:

- Energy window: 10 eV (`-e 10`), capturing all relevant S$_1$–S$_n$ and T$_1$–T$_n$ states
- Configuration selection: automated based on energy threshold
- sTDA: Tamm-Dancoff approximation (faster, typically sufficient for absorption spectra)
- sTD-DFT: Full RPA coupling (more accurate for emission properties)

S1.4 NTO analysis with Multiwfn

Natural Transition Orbitals were generated using Multiwfn with the following workflow:

1. Load excited-state molden file
2. Main function 18 (Electron excitation analysis)
3. Subfunction 1 (NTO analysis)
4. Export NTO pairs for hole and electron

The hole-electron overlap S_{he} was computed as:

$$S_{he} = \sum_A \sqrt{\rho_h^A \cdot \rho_e^A} \tag{S1}$$

where ρ_h^A and ρ_e^A are the Mulliken populations of hole and electron on atom A.

S2 TADF Photophysics: Theoretical Background

S2.1 Singlet-triplet energy gap

The singlet-triplet energy gap is defined as:

$$\Delta E_{\mathrm{ST}} = E(S_1) - E(T_1) \tag{S2}$$

For efficient TADF, $\Delta E_{\mathrm{ST}} < 0.2$ eV is required to enable thermal upconversion at room temperature ($k_B T \approx 0.026$ eV).

S2.2 Reverse intersystem crossing rate

The RISC rate follows Marcus-type kinetics:

$$k_{\mathrm{RISC}} = \frac{2\pi}{\hbar} |\langle S_1 | \hat{H}_{\mathrm{SOC}} | T_1 \rangle|^2 \frac{1}{\sqrt{4\pi\lambda k_B T}} \exp\left(-\frac{(\Delta E_{\mathrm{ST}} + \lambda)^2}{4\lambda k_B T}\right) \tag{S3}$$

where:

- $\langle S_1 | \hat{H}_{\mathrm{SOC}} | T_1 \rangle$ is the spin-orbit coupling matrix element
- λ is the reorganization energy ($\sim$0.1 eV for rigid systems)
- $k_B T$ is the thermal energy (0.026 eV at 300 K)

S2.3 El-Sayed's rules

According to El-Sayed's rules, SOC is maximized when the transition involves a change in orbital character:

$$|\langle S_1 | \hat{H}_{\mathrm{SOC}} | T_1 \rangle| \propto |\Delta \mathrm{Character}_{S_1 - T_1}| \tag{S4}$$

For TADF emitters, this translates to:

- S_1: predominantly ^{1}CT (charge-transfer) character
- T_1: mixed ^{3}CT/^{3}LE (local excitation) character

The NTO overlap difference $\Delta S_{\text{NTO}} = S_{\text{NTO}}^{T_1} - S_{\text{NTO}}^{S_1}$ quantifies this character difference.

S3 High-Throughput Screening Protocol

S3.1 Dataset composition

The 747-molecule dataset comprises TADF emitters from four architectural classes:

Table S1 Dataset composition by molecular architecture

Architecture	Count	Percentage
D–A (Donor–Acceptor)	198	26.5%
D–A–D	312	41.8%
MR (Multi-Resonance)	89	11.9%
TSCT (Through-Space CT)	148	19.8%
Total	747	100%

S3.2 Screening workflow

The hierarchical screening protocol consists of:

1. **Structure preparation**: SMILES → 3D coordinates (RDKit)
2. **Conformer search**: CREST with GFN2-xTB
3. **Geometry optimization**: GFN2-xTB (gas and toluene)
4. **Excited states**: sTDA/sTD-DFT-xTB
5. **NTO analysis**: Multiwfn
6. **CT descriptor extraction**: Custom Python scripts

S3.3 Property definitions

Key properties extracted from calculations:

- E_{S_1}: Vertical S_1 excitation energy (eV)
- E_{T_1}: Vertical T_1 excitation energy (eV)
- ΔE_{ST}: $E_{S_1} - E_{T_1}$ (eV)
- f_{S_1}: Oscillator strength of S_1 transition
- E_{gap}: HOMO-LUMO gap (eV)

S4 NTO and CT Descriptor Definitions

S4.1 Natural Transition Orbitals

NTOs provide a compact representation of electronic transitions. For a transition from ground state $|0\rangle$ to excited state $|n\rangle$, the transition density matrix is:

$$\gamma_{pq}^{0n} = \langle n|\hat{a}_p^\dagger \hat{a}_q|0\rangle \tag{S5}$$

Singular value decomposition yields hole (ϕ_i^h) and particle (ϕ_i^e) NTOs:

$$\gamma^{0n} = \mathbf{U}\Sigma\mathbf{V}^\dagger \tag{S6}$$

S4.2 Charge-transfer descriptors

The following CT descriptors were computed from NTO analysis:

Table S2 CT descriptor definitions

Descriptor	Definition
S_{he}	Hole-electron spatial overlap: $\sum_A \sqrt{\rho_h^A \cdot \rho_e^A}$
Ω_{CT}	CT number: fraction of transition with CT character
Λ_D	Hole localization on donor fragment
Λ_A	Electron localization on acceptor fragment
Δr	Hole-electron centroid distance (Å)
S_{NTO}	NTO orbital overlap

S4.3 Physical interpretation

The hole-electron overlap S_{he} directly relates to the exchange interaction:

$$K_{ij} \propto \iint |\phi_h(\mathbf{r}_1)|^2 \frac{1}{r_{12}} |\phi_e(\mathbf{r}_2)|^2 d\mathbf{r}_1 d\mathbf{r}_2 \tag{S7}$$

Low S_{he} indicates spatial separation of hole and electron, leading to:

- Reduced exchange interaction
- Smaller ΔE_{ST}
- Enhanced CT character

S5 Active Learning: Acquisition Functions

S5.1 Uncertainty Sampling (US)

Selects samples with highest predictive uncertainty:

$$\alpha_{\text{US}}(x) = \sigma(x) \tag{S8}$$

where $\sigma(x)$ is the standard deviation of predictions across Random Forest trees.

S5.2 Expected Improvement (EI)

Balances exploration and exploitation for minimization:

$$\alpha_{\mathrm{EI}}(x) = (\mu^* - \mu(x))\Phi(Z) + \sigma(x)\phi(Z) \tag{S9}$$

where $Z = \frac{\mu^* - \mu(x)}{\sigma(x)}$, μ^* is the best observed value.

S5.3 Upper Confidence Bound (UCB)

Optimistic acquisition for minimization:

$$\alpha_{\mathrm{UCB}}(x) = -\mu(x) + \kappa\sigma(x) \tag{S10}$$

with exploration parameter $\kappa = 2.0$.

S5.4 Query by Committee (QBC)

Uses disagreement among ensemble members:

$$\alpha_{\mathrm{QBC}}(x) = \frac{1}{C}\sum_{c=1}^{C}(f_c(x) - \bar{f}(x))^2 \tag{S11}$$

where C is the committee size (10 models).

S5.5 Diversity Sampling

Maximizes distance to existing training samples:

$$\alpha_{\mathrm{div}}(x) = \min_{x' \in \mathscr{D}_{\mathrm{train}}} \|x - x'\|_2 \tag{S12}$$

S5.6 Hybrid (Uncertainty × Diversity)

Combines uncertainty and diversity:

$$\alpha_{\mathrm{hybrid}}(x) = \alpha \cdot \tilde{\sigma}(x) + (1 - \alpha) \cdot \tilde{d}(x) \tag{S13}$$

where $\tilde{\sigma}$ and $\tilde{d}$ are normalized scores, $\alpha = 0.5$.

S6 Supplementary Tables

Table S3 Hyperparameter grid search for ML models

Model	Parameter	Search Range
RF	n_estimators	[100, 200, 500]
	max_depth	[10, 20, None]
	min_samples_split	[2, 5, 10]
SVR	C	[0.1, 1, 10, 100]
	gamma	[0.01, 0.1, 1, auto]
	kernel	[rbf]
GB	n_estimators	[100, 200, 500]
	learning_rate	[0.01, 0.1, 0.2]
	max_depth	[3, 5, 7]

Table S4 High-level theory validation results

Molecule	sTD-DFT-xTB	HLT Ref.	Error
4CzIPN	0.21 eV	0.16 eV	0.05 eV
DMAC-TRZ	0.085 eV	0.05 eV	0.035 eV
MAE			**0.045 eV**

Table S5 Feature importance by category (SHAP analysis)

Category	Features	Importance
Energy	E_{T_1}, E_{S_1}, E_{gap}	57%
CT descriptors	$S_{he}^{T_1}$, $S_{he}^{S_1}$, Δr, Λ	34%
Oscillator strength	f_{S_1}	8%
NTO overlap	$S_{\mathrm{NTO}}^{S_1}$, $S_{\mathrm{NTO}}^{T_1}$	1%

S7 OT-LC-PBE Validation Calculations

To validate the xTB-based protocol with explicit high-level calculations, we performed optimally-tuned long-range corrected PBE (OT-LC-PBE) calculations on three representative TADF molecules spanning different architectures and sizes.

S7.1 Computational methodology

All OT-LC-PBE calculations were performed using ORCA 6.1.0. The protocol consists of two steps:

*S7.1.0.1 Optimal ω tuning*The range-separation parameter ω was optimized by minimizing the IP-tuning criterion:

$$J(\omega) = |\varepsilon_{\mathrm{HOMO}}(\omega) + IP(\omega)| \tag{S14}$$

where $\varepsilon_{\mathrm{HOMO}}(\omega)$ is the HOMO eigenvalue and $IP(\omega) = E(N-1;\omega) - E(N;\omega)$ is the vertical ionization potential. Calculations used the def2-SVP basis set with LC-PBE functional. Initial grid search (11 points, $\omega \in [0.10, 0.30]$ bohr^{-1}) was followed by golden-section refinement to $J < 10^{-4}$ Ha.

*S7.1.0.2 TD-DFT excited states*Vertical excitation energies were computed using full TD-DFT (not TDA) at the optimized ω values with:

- Functional: LC-PBE with molecule-specific ω
- Basis set: def2-TZVP
- Auxiliary basis: def2/J with RIJCOSX approximation
- States: 10 singlets, 10 triplets
- Geometry: GFN2-xTB optimized (same as HTS protocol)
- Parallelization: 8 cores, 2.5 GB/core

Table S6 Top TADF candidates for each application

Application	Molecule	ΔE_{ST}	Key Property
Bioimaging	PXZ-NAI	0.29 eV	$\lambda_{\mathrm{em}} \approx 690$ nm
Photocatalysis	TPA-APy	-0.034 eV	Inverted gap
Photodetection	BMZ-TZ	-0.006 eV	$S_{he}^{T_1} = 0.89$

Table S7 OT-LC-PBE results for benchmark TADF molecules. All energies in eV.

Molecule	Atoms	Arch.	ω (bohr^{-1})	E_{S_1}	E_{T_1}	ΔE_{ST}
BACN	48	A-D-A	0.181	3.26	2.46	0.81
DMAC-TRZ	68	D-A	0.185	3.10	2.93	0.17
4CzIPN	94	4D-A	0.147	2.69	2.49	0.19

S7.2 Results

S7.3 Key observations

1. **Optimal ω correlates with CT character**: 4CzIPN (strongest CT) has the smallest ω (0.147 bohr^{-1}), consistent with more delocalized frontier orbitals.

2. **Vertical vs. adiabatic discrepancy**: OT-LC-PBE vertical TD-DFT overestimates ΔE_{ST} compared to experimental (adiabatic) values. For DMAC-TRZ: 0.17 eV (vertical) vs. 0.05 eV (exp.).

3. **xTB provides conservative estimates**: The xTB methods systematically underestimate ΔE_{ST} by 0.08–0.35 eV relative to OT-LC-PBE. This is advantageous for screening, as molecules passing the threshold will have true ΔE_{ST} at least as favorable.

4. **Ranking preserved**: Despite systematic offsets, the relative ordering of molecules by ΔE_{ST} is maintained across methods, validating xTB for prioritization in high-throughput workflows.

5. **Computational cost**: BACN ($\sim$1.5 h), DMAC-TRZ ($\sim$2.5 h), 4CzIPN ($\sim$9 h) on 8 cores, compared to $<$1 min for xTB calculations.

S8 Supplementary Figures

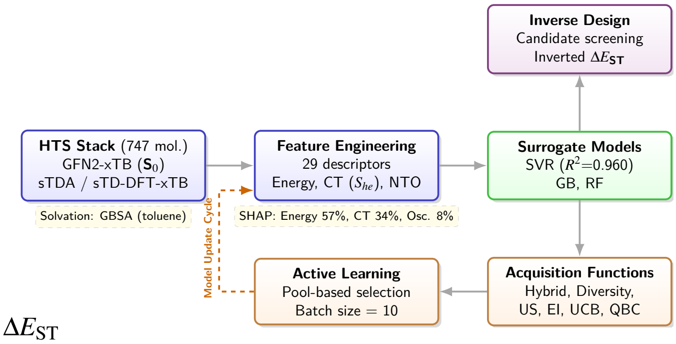

Fig. S1 Schematic overview of the ML/Active Learning workflow. The pipeline integrates semi-empirical quantum chemistry (GFN2-xTB, sTDA) with NTO-based feature extraction, surrogate model training, and iterative active learning for efficient TADF property prediction.

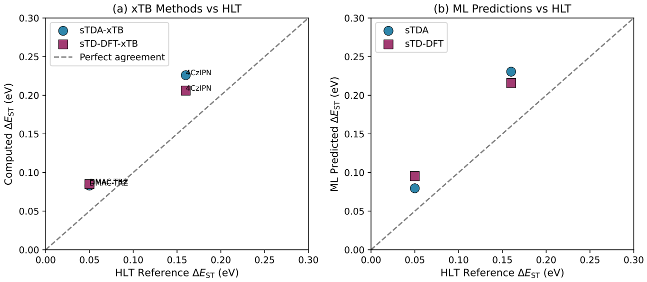

Fig. S2 Validation of sTD-DFT-xTB predictions against high-level theory (HLT) references from the STGABS27 benchmark set. The MAE of 0.045 eV demonstrates chemical accuracy suitable for ML-driven discovery.

S9 Data Availability and Reproducibility

S9.1 Dataset files

The following files are available at Zenodo (DOI: 10.5281/zenodo.17436069):

- `combined_features_747mol.csv` – Full feature table (2943 samples × 42 features)
- `nto_orbital_overlap_747mol.csv` – NTO overlap data
- `ct_descriptors_747mol.csv` – CT descriptor values
- `ml_results_747mol.json` – Model performance metrics
- `al_results_747mol.json` – Active learning results
- `predictions_747mol.csv` – Model predictions

S9.2 Code repository

Python scripts for reproducing all results:

- `ml_pipeline_747mol.py` – ML model training
- `al_experiment_747mol.py` – Active learning experiments

- `generate_figures_747mol.py` – Figure generation

- `compute_ct_descriptors_747mol.py` – CT descriptor extraction

S9.3 Environment

A complete environment specification (`requirements.txt`) is provided:

```
numpy==2.0.2
pandas==2.2.3
scikit-learn==1.7.2
shap==0.50.0
matplotlib==3.9.3
scipy==1.14.1
```