Supporting Information

Extracting Recalcitrant Redox Data on Fluorophores to Pair with Optical Data for Predicting Small-Molecule, Ionic Isolation Lattices

Michaela K. Loveless[†a], Minwei Che[†a], Alec J. Sanchez[a], Vikrant Tripathy[a], Bo W. Laursen[b], Sudhakar Pamidighantam[a,c,d], Krishnan Raghavachari[a], Amar H. Flood[a]*

a. Department of Chemistry, Indiana University 800 E. Kirkwood Avenue, Bloomington, IN 47405, USA.

b. Nano-Science Center & Department of Chemistry, University of Copenhagen, Universitetsparken 5, 2100 Copenhagen, Denmark

c. Center for AI in Science and Engineering (ARTISAN), Georgia Institute of Technology 1283B CODA Building, 756 W Peachtree St NW, Atlanta, GA 30308, USA.

d. Institute for Data Engineering and Science (IDEaS), Georgia Institute of Technology 1283B CODA Building, 756 W Peachtree St NW, Atlanta, GA 30308, USA.

Table of Contents

Methods of Natural Language Processing

The use of natural language processing (NLP) in this work is limited to the extraction of optical data using ChemDataExtractor (CDE).[1,2] Much of this extraction was originally published at the Deep4Chem. Optical data of additional fluorophores were scraped with CDE while electrochemical data were collected by hand. Manual extraction of electrochemical data was necessary due to the data's variety in experimental details and reporting practices. Further discussion on the strengths and weaknesses of these methods as well as the outcome of these methods can be found in the main text.

Additional Optical vs. Redox Gap Plots

(a)
y = 1.193x - 0.655
R² = 0.773
n = 9

(b)
y = 1.288x - 0.723
R² = 0.511
n = 8

(c)
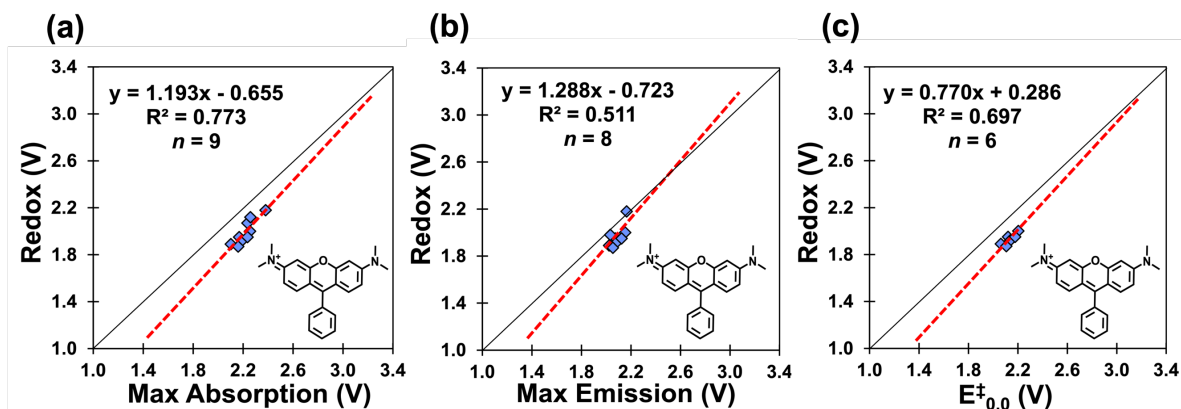y = 0.770x + 0.286
R² = 0.697
n = 6

Figure S1. Plots of electrochemical window between oxidation and reduction versus (a) absorption maxima, (b) emission maxima and (c) the estimates for the $E_{0,0}^{\ddagger}$ defined as the numerical average of the absorption and emission maxima for **rhodamines**.
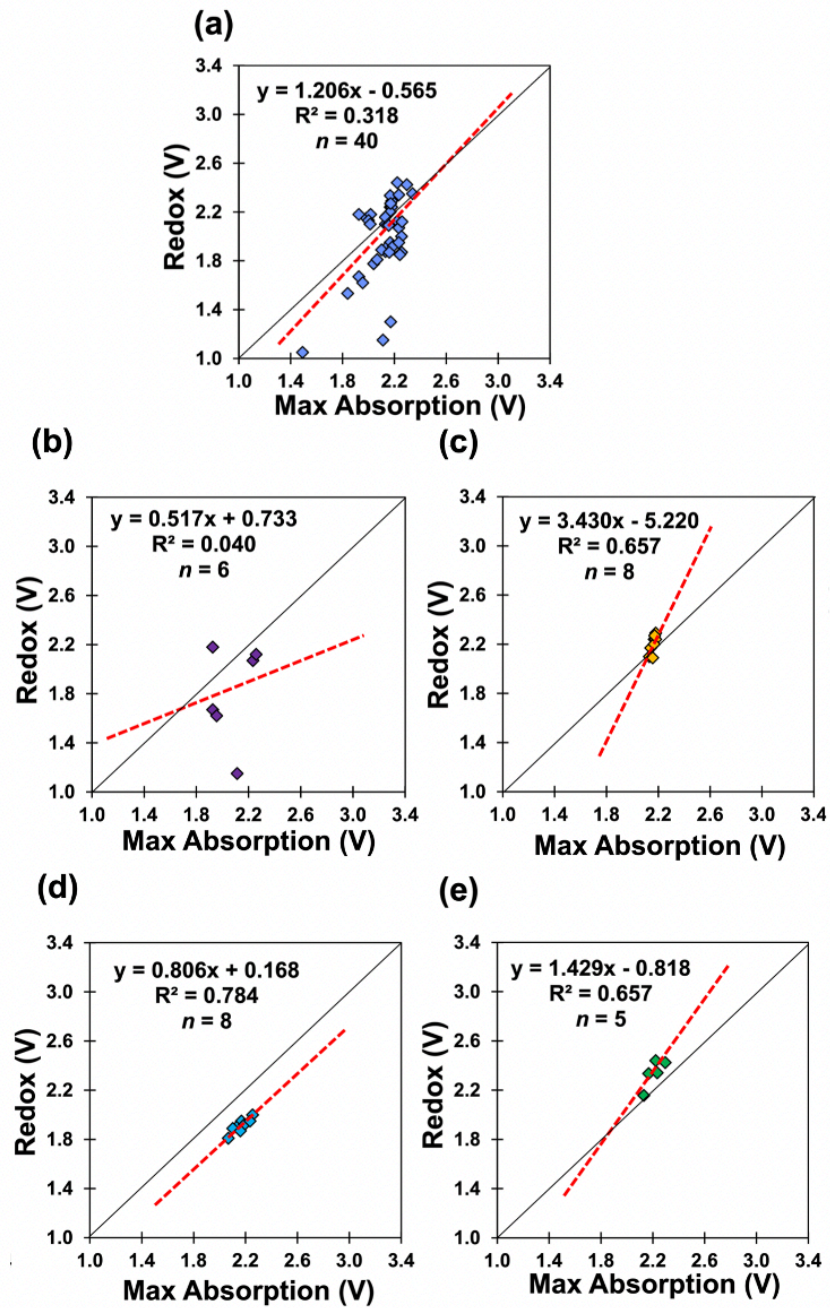
Figure S2. Plots of electrochemical window between oxidation and reduction versus absorption maxima for the (a) full dataset ($n = 40$ datapoints) and for dyes within a single publication (b-e) with $n > 4$ datapoints. Colors are used to distinguish between the different papers.
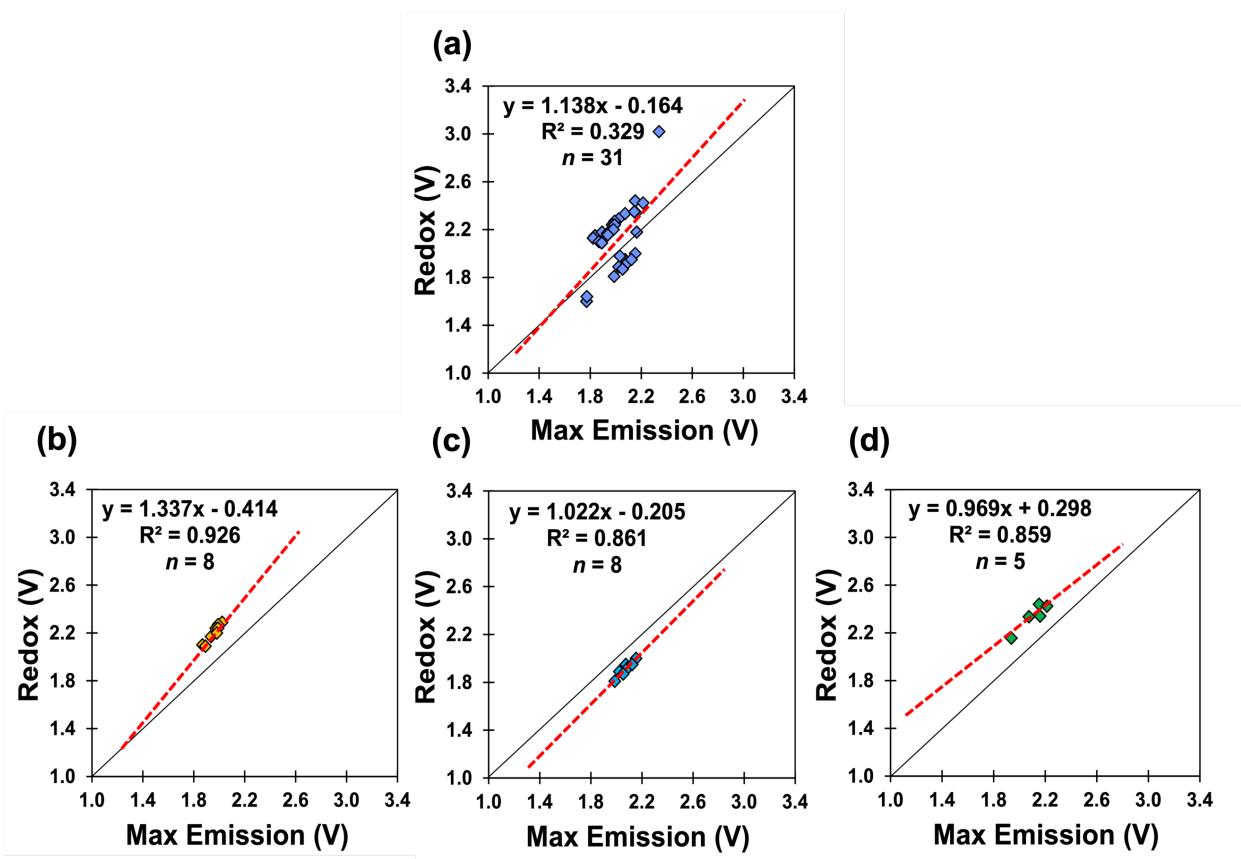
(a)

(b)

(c)

(d)

Figure S3. Plots of electrochemical window between oxidation and reduction versus emission maxima for the (a) full dataset ($n = 31$ datapoints) and for dyes within a single publication (b-d) with $n > 4$ datapoints. Colors are used to distinguish between the different papers.

(a)

$$y = 0.484x + 1.127$$
$$R^2 = 0.065$$
$$n = 26$$

(b)

$$y = 2.022x - 1.962$$
$$R^2 = 0.893$$
$$n = 8$$

(c)

$$y = 0.914x - 0.024$$
$$R^2 = 0.830$$
$$n = 8$$

(d)

$$y = 1.189x - 0.227$$
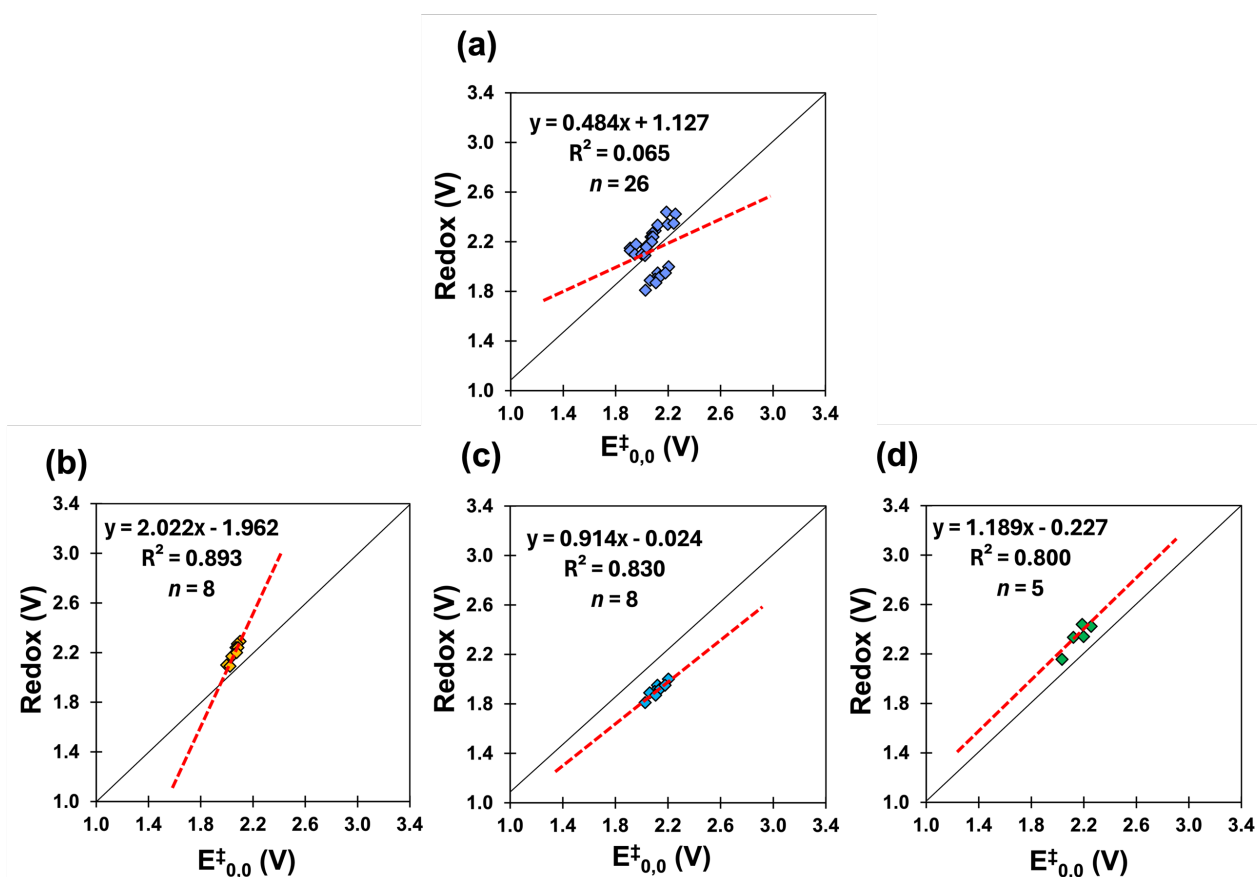$$R^2 = 0.800$$
$$n = 5$$

Figure S4. Plots of electrochemical window between oxidation and reduction versus the estimates for the $E_{0,0}^{\ddagger}$ defined as the numerical average of the absorption and emission maxima for the (a) full dataset ($n = 26$ datapoints) and for dyes within a single publication (b-d) with $n > 4$ datapoints. Colors are used to distinguish between the different papers.

From plots of redox versus optical gaps (Figure 12a-c), it is evident that the data is weakly correlated ($0.065 \leq R^2 \leq 0.329$), with most of the data falling in two regions that have their own trends. These two regions are dominated by two dye families, the trianguleniums and rhodamines. When plotting the electrochemical and optical data based on dye family (Figure 12e-f and FigureS1a-c), the correlation increases substantially ($0.511 \leq R^2 \leq 0.912$). Similarly, when dyes are grouped based on a single paper, the correlation improves drastically ($R^2 \leq 0.625$, Figures S2-S4) aside from one which has opposite trends (Figure S2a). These observations suggest that electrochemical and optical data are sensitive to dye family and experimental conditions, which are likely to be self-consistent across a single paper and which we posit to be idiosyncratic to a single experimentalist. This finding highlights the need for systemic procedures when reporting electrochemical data, such as those recently recommended by the editors of the *American Chemical Society*.[3]

References

1. J. Mavracic, C. J. Court, T. Isazawa, S. R. Elliott and J. M. Cole, *J. Chem Inf. Model.*, 2021, **61**, 4280-4289.
2. M. C. Swain and J. M. Cole, *J. Chem Inf. Model.*, 2016, **56**, 1894-1904.
3. S. Minteer, J. G. Chen, S. Lin, C. Crudden, S. Dehnen, P. V. Kamat, M. Kozlowski, G. Masson and S. J. Miller, *Journal of Organic Chemistry*, 2023, **88**, 4036-4037.