Electronic supplementary information

Data-Driven Analysis of Hole-Transporting Materials for Perovskite Solar Cells Performance

Marcos del Cueto*¹, Charles Rawski-Furman¹, Juan Aragó², Enrique Ortí², Alessandro Troisi³

1. Department of Chemistry, University of Liverpool, Liverpool, L69 3BX, UK
2. Instituto de Ciencia Molecular (ICMol), Universidad de Valencia, Catedrático José Beltrán 2, Paterna 46980, Spain

Contents:

S1. HTM features
   a. Structural features
   b. Electronic features
S2. Conformer search
S3. ML model
   a. Differential evolution algorithm
   b. Heterogeneous database parameters
   c. Homogeneous database parameters
S4. kNN results
S5. Chemical fragments correlation with PCE
S6. Database
S7. References
S1. HTM features

a. Structural features

We initially consider the following 32 structural features:

1. Phthalocyanine
2. Porphyrin
3. Azulene
4. Spiro atoms
5. Thiophene
6. Furan
7. Fluorene
8. Carbazole
9. Dibenzofuran
10. Diphenylamine
11. Triphenylamine

12. Polymer (boolean)
13. Xanthene
14. Acenaphthene
15. Benzotriphenylene
16. Silicon atoms
17. Rotatable bonds
18. Molecular weight
19. sp³ Carbons
20. Aliphatic carbocycles
21. Aliphatic heterocycles
22. Aliphatic rings

23. Amide bonds
24. Aromatic carbocycles
25. Aromatic heterocycles
26. Aromatic rings
27. Stereocenters
28. Bridgehead atoms
29. Heterocycles
30. Heteroatoms
31. Rings
32. Molecular planarity (measured as the sum of atomic distances from plane of best fit - PBF)¹

The correlation matrix of all these features are shown in Figure S1 (for the heterogeneous database), and we then remove those features with a Pearson correlation larger than 0.7:

- molecular weight (large correlation with rotatable bonds)
- fluorenes (large correlation with aliphatic carbocycles)
- aliphatic rings (large correlation with aliphatic heterocycles)
- thiophenes (large correlation with aromatic heterocycles)
- aromatic rings (large correlation with aromatic carbocycles)
- rings (large correlation with aromatic carbocycles)
- heterocycles (large correlation with aromatic heterocycles)
- bridgehead atoms (large correlation with phthalocyanines)

Figure S1. Correlation matrix with the initial 32 structural features in the heterogeneous dataset.
Finally, we used a $k$-nearest neighbors model, using all the 269 data points to estimate the root mean square error (rmse) when using a 10-fold cross-validation. We performed a recursive feature selection, whose results are shown in Figure S2, and the following descriptors are dropped sequentially: spiro atoms, amide bonds, azulene, diphenylamine, xanthene, aromatic carbocycles, heteroatoms, aromatic heterocycles, molecular planarity, sp$^3$ carbons, furan, dibenzofuran, porphyrin, aliphatic carbocycle, stereocenters, silicon atoms, phthalocyanine, benzotri thiophene, triphenylamine, acenaphthene, aliphatic heterocycles, rotatable bonds, carbazole. The lowest rmse is achieved when using 12 structural features, corresponding to:

1. Phthalocyanine
2. Porphyrin
3. Carbazole
4. Triphenylamine
5. Polymer
6. Acenaphthene
7. Benzo tri thiophene
8. Silicon atoms
9. Rotatable bonds
10. Aliphatic carbocycles
11. Aliphatic heterocycles
12. Stereocenters

Figure S2. rmse values obtained with a feature recursive elimination in the heterogeneous database.

In the case of the homogeneous database, there is only one molecule with a phthalocyanine group, and there are no porphyrins present, so we do not consider these two features. In Figure S3, we show the resulting correlation matrix with the remaining 30 structural features. We remove the following nine features with a Pearson correlation larger than 0.7:

- Dibenzofuran (large correlation with furan)
- Molecular weight (large correlation with rotatable bonds)
- Fluorene (large correlation with aliphatic carbocycles)
- Azulene (large correlation with aliphatic heterocycles)
- Aliphatic rings (large correlation with aliphatic carbocycles and aliphatic heterocycles)
- Thiophene (large correlation with heterocycles and aromatic heterocycles)
- Aromatic rings (large correlation with aromatic carbocycles)
- Heterocycles (large correlation with aromatic heterocycles)
- Rings (large correlation with aromatic carbocycles)
We use the same procedure as in the heterogeneous database to perform a recursive feature selection on the remaining 21 structural features of the homogeneous database, as shown in Figure S4, where we drop the following features sequentially: diphenylamine, amide bonds, carbazole, aromatic carbocycles, bridgehead atoms, silicon atoms, xanthene, furan, molecular planarity, polymer, aromatic heterocycles, aliphatic heterocycles, stereocenters, heteroatoms, aliphatic carbocycles, benzothiophene, rotatable bonds, spiro atoms, triphenylamine. We can observe how the optimum number of structural features is reached with nine features:

1. Spiro atoms
2. Carbazole
3. Triphenylamine
4. Acenaphthene
5. Benzothiophene
6. sp² Carbons
7. Aliphatic heterocycles
8. Aromatic heterocycles
9. Molecular planarity (PBF)
b. Electronic features

![Figure S5](image)

**Figure S5.** Ionization potential (IP) and HOMO energy data in the heterogeneous database.

### S2. Conformer search

Starting from the SMILES string of each HTM, we calculated the ensemble of the ten most stable conformers with a UFF forcefield, using Open Babel. Then, we optimized these ten geometries using PM7 with Gaussian16, and the most stable geometry was used as a starting point for the subsequent density functional theory (DFT) calculations.

### S3. ML model

##### a. Differential evolution algorithm

We used a differential evolution algorithm, as implemented in SciPy, using a population size of 15 per parameter, a recombination rate of 0.7 and a mutation of 0.5-1.0.

##### b. Heterogeneous database parameters

**Table S1.** Optimized parameters and resulting rmse and correlation coefficient (r) values with the heterogeneous database (values within parentheses correspond to the test set, and those outside of parentheses correspond to the training set).

| $\gamma_{fam}$ | $\gamma_{arch}$ | $\gamma_{fp}$ | $\gamma_{str}$ | $\gamma_{elec}$ | $\gamma_{add}$ | $\alpha$ | rmse       | r           |
|-----------------|-----------------|---------------|----------------|-----------------|----------------|---------|------------|-------------|
| 0.179072        | 0.154827        | 1.0           | 0.000012       | 0.000640        | 0.020764       | 0.431047 | 3.0% (3.0%)| 0.68 (0.72) |
c. Homogeneous database parameters

Table S2. Optimized parameters and resulting rmse and r values with the homogeneous database (values within parentheses correspond to the test set, and those outside of parentheses correspond to the training set).

| Type                        | $\gamma_{fp}$ | $\gamma_{str}$ | $\gamma_{elec}$ | $\gamma_{add}$ | $\alpha$ | rmse   | r      |
|-----------------------------|----------------|-----------------|------------------|----------------|----------|--------|--------|
| Fingerprint + Additives     | 1.0            | 0               | 0                | 0.129438       | 0.042468  | 2.8% (2.8%) | 0.59 (0.54) |
| Structural + Additives      | 0              | 1.0             | 0                | 0.030761       | 0.004050  | 6.9% (3.9%) | 0.19 (0.27) |
| Electronic + Additives      | 0              | 0               | 1.0              | 0.151269       | 0.022335  | 4.1% (5.4%) | 0.40 (0.65) |
| Fingerprint + Structural + Electronic + Additives | 1.0 | 0.000155 | 0.000026 | 0.071131 | 0.040924 | 2.8% (2.7%) | 0.59 (0.57) |

Figure S6. Experimental and predicted PCE of data in the homogeneous database, when using different types of features.
S4. kNN results

Using the $k$-nearest neighbours algorithm, we obtain the following results with the homogeneous dataset, which present a similar trend to the KRR results.

**Table S3.** Optimized parameters and resulting $rmse$ and $r$ values with the homogeneous database using kNN (values within parentheses correspond to the test set, and those outside of parentheses correspond to the training set).

| Type | $γ_{fp}$ | $γ_{str}$ | $γ_{elec}$ | $γ_{add}$ | $k$ | $rmse$ | $r$ |
|------|----------|-----------|------------|-----------|-----|--------|-----|
| Fingerprint + Structural + Electronic + Additives | 7.637 | 7.403 | 0.677 | 2.664 | 3 | 2.4% (2.1%) | 0.73 (0.77) |

![Figure S7. Experimental and predicted PCE of data in the homogeneous database using kNN, when using different types of features.](image)

S5. Chemical fragments correlation with PCE

Given than most of the model performance is due to the fingerprints, we can analyse which bits of the fingerprints are more correlated with PCE.

For each bit, we created an array of length $N$ (where $N$ is the number of molecules in the dataset), which will have values of either 0 (bit absent in that molecule) or 1 (bit present in that molecule). Then, we can study the correlation of these with an array containing the PCE values of each molecule in the dataset. We used a point biserial correlation coefficient ($r$), which is equivalent to Pearson’s correlation when one variable is binary. We give this value, its corresponding p-value (as the probability of observing the same or larger $|r|$ if data is uncorrelated) and the 95% confidence interval below, for the 10 fragments with the largest $|r|$ values.

Additionally, we have performed a Mann-Whitney U test for each fragment, which measures the statistical difference between two arrays: $PCE_1$ and $PCE_0$, where $PCE_1$ contains the PCE values of the molecules with that fragment present and $PCE_0$ contains the PCE values of the molecules with that fragment absent. The Mann-Whitney U statistics is defined as:

$$U = \sum_{i=1}^{n} \sum_{j=1}^{m} S(X_i, Y_i)$$

with $S(X, Y) = \begin{cases} 1 & \text{if } X > Y \\ 1/2 & \text{if } X = Y \\ 0 & \text{if } X < Y \end{cases}$. We report in the table below the $U_1$ and $U_2$ statistics, where $U_1 = \sum_{i=1}^{n} \sum_{j=1}^{m} S(PCE_{1,i}, PCE_{0,j})$ and $U_2 = \sum_{i=1}^{n} \sum_{j=1}^{m} S(PCE_{0,i}, PCE_{1,j})$. For example, a large value of $U_1$ indicates that there is a significant difference between the $PCE_0$ and $PCE_1$ arrays, with $PCE_1$ having larger values.
The results for the 19 molecules with $|r| > 0.3$ in the homogeneous database are shown in Table S3, and the results for the 27 molecules with $|r| > 0.2$ in the heterogeneous database are shown in Table S4.

**Table S4.** Results for the bits whose correlation coefficient $|r| > 0.3$ with respect to PCE for all molecules in the homogeneous dataset, with its corresponding p-value and 95% confidence interval, as well as the Mann-Whitney U test values ($U_1$ and $U_2$) and its associated p-value for each fragment.

| Bit | r   | p-value     | $r$ 95% interval | $U_1$ | $U_2$ | p-value   |
|-----|-----|-------------|------------------|-------|-------|-----------|
| 832 | 0.47| 1.08E-06    | [0.30,0.61]      | 1919.5| 576.5 | 3.65E-06  |
| 842 | 0.42| 1.25E-05    | [0.25,0.57]      | 1837.5| 637.5 | 3.26E-05  |
| 1537| 0.42| 1.25E-05    | [0.25,0.57]      | 1837.5| 637.5 | 3.26E-05  |
| 719 | 0.41| 2.53E-05    | [0.23,0.56]      | 1807.5| 656.5 | 6.51E-05  |
| 15  | -0.41|2.85E-05     | [-0.56,-0.23]    | 162   | 817   | 3.15E-04  |
| 323 | 0.40| 3.03E-05    | [0.23,0.56]      | 1792.5| 658.5 | 7.99E-05  |
| 171 | 0.40| 4.31E-05    | [0.22,0.55]      | 1670.5| 604.5 | 1.19E-04  |
| 882 | 0.40| 4.31E-05    | [0.22,0.55]      | 1670.5| 604.5 | 1.19E-04  |
| ID   | X    | Y    | Z    | Item 1 | Item 2 | Item 3 | Item 4 | Item 5 | Item 6 |
|------|------|------|------|--------|--------|--------|--------|--------|--------|
| 961  | 0.40 | 4.31E-05 | [0.22,0.55] | 1670.5 | 604.5 | 1.19E-04 |
| 1097 | 0.40 | 4.31E-05 | [0.22,0.55] | 1670.5 | 604.5 | 1.19E-04 |
| 1568 | -0.38 | 8.35E-05 | [-0.54,-0.20] | 207.5 | 848.5 | 0.000687 |
| 782  | 0.37 | 1.32E-04 | [0.19,0.53] | 1751.5 | 699.5 | 0.000253 |
| 2033 | 0.35 | 3.83E-04 | [0.16,0.51] | 1686.5 | 732.5 | 0.000838 |
| 740  | -0.34 | 4.95E-04 | [-0.50,-0.16] | 336 | 1008 | 0.001605 |
| 1212 | -0.33 | 7.90E-04 | [-0.49,-0.14] | 438 | 1162 | 0.001836 |
| 123  | -0.33 | 8.55E-04 | [-0.49,-0.14] | 135.5 | 515.5 | 0.010454 |
| 1445 | -0.31 | 1.51E-03 | [-0.48,-0.12] | 693 | 1582 | 0.001331 |
| 1936 | -0.30 | 2.28E-03 | [-0.47,-0.11] | 88 | 476 | 0.004971 |
Table S5. Results for the bits whose correlation coefficient $|r| > 0.2$ with respect to PCE for all molecules in the heterogeneous dataset, with its corresponding p-value and 95% confidence interval, as well as the Mann-Whitney U test values ($U_1$ and $U_2$) and its associated p-value for each fragment.

| Bit | $r$ | p-value  | $r$ 95% interval | $U_1$ | $U_2$ | p-value |
|-----|-----|----------|------------------|------|------|--------|
| 1097| 0.32| 1.24E-07 | [0.20,0.42]     | 10077| 4821 | 5.67E-06|
| 882 | 0.30| 6.73E-07 | [0.18,0.40]     | 9993.5| 5126.5| 3.03E-05|
| 842 | 0.27| 8.14E-06 | [0.15,0.38]     | 11279| 6211 | 5.37E-05|
| 1936| -0.27|9.58E-06 | [-0.37,-0.15]   | 125 | 1453 | 4.29E-04|
| 171 | 0.27| 1.03E-05 | [0.15,0.37]     | 9875 | 5665 | 3.72E-04|
| 961 | 0.26| 1.32E-05 | [0.15,0.37]     | 9906 | 5634 | 3.04E-04|
| 1537| 0.26| 1.94E-05 | [0.14,0.37]     | 11173.5| 6364.5| 1.30E-04|
| 832 | 0.25| 2.48E-05 | [0.14,0.36]     | 11256.5| 6677.5| 3.14E-04|
|   |   |   |   |   |   |   |
|---|---|---|---|---|---|---|
| 2033 | 0.24 | 5.21E-05 | [0.13,0.35] | 10357 | 6251 | 7.86 E-04 |
| 323  | 0.24 | 8.19E-05 | [0.12,0.35] | 10665 | 6495 | 7.94 E-04 |
| 1446 | -0.24 | 8.47E-05 | [-0.35,0.12] | 609 | 2475 | 4.00E-04 |
| 719  | 0.23 | 1.24E-04 | [0.12,0.34] | 10691.5 | 6642.5 | 1.19E-03 |
| 2    | -0.22 | 2.45E-04 | [-0.33,-0.11] | 3266 | 6082 | 2.15E-03 |
| 740  | -0.22 | 2.62E-04 | [-0.33,0.11] | 2744.5 | 5245.5 | 3.20E-03 |
| 335  | -0.22 | 2.92E-04 | [-0.33,-0.10] | 178 | 1400 | 1.20E-03 |
| 9    | -0.22 | 2.99E-04 | [-0.33,-0.10] | 94 | 966 | 4.80E-03 |
| 1092 | -0.22 | 2.99E-04 | [-0.33,-0.10] | 94 | 966 | 4.80E-03 |
| 1094 | -0.22 | 2.99E-04 | [-0.33,-0.10] | 94 | 966 | 4.80E-03 |
|   |   |   |   |   |   |   |
|---|---|---|---|---|---|---|
| **1273** | -0.22 | 2.99E-04 | [-0.33,-0.10] | 94 | 966 | 4.80E-03 |
| **1815** | -0.22 | 2.99E-04 | [-0.33,-0.10] | 94 | 966 | 4.80E-03 |
| **364**  | -0.22 | 3.82E-04 | [-0.33,-0.10] | 4  | 530 | 1.66E-02 |
| **1552** | -0.22 | 3.82E-04 | [-0.33,-0.10] | 4  | 530 | 1.66E-02 |
| **2008** | -0.21 | 5.71E-04 | [-0.32,-0.09] | 8  | 526 | 1.84E-02 |
| **1172** | -0.21 | 6.80E-04 | [-0.32,-0.09] | 94.5 | 965.5 | 4.85E-03 |
| **592**  | -0.20 | 9.07E-04 | [-0.31,-0.08] | 6542 | 10288 | 2.34E-03 |
| **1912** | -0.20 | 9.07E-04 | [-0.31,-0.08] | 6542 | 10288 | 2.34E-03 |
| **758**  | -0.20 | 9.60E-04 | [-0.31,-0.09] | 11.5 | 522.5 | 2.00E-02 |
S6. Database
Figure S8. Structure of all HTMs in our database. Labels include the PCE and perovskite family/architecture of each device. In case of polymeric HTMs, we show the structure of the repeating unit.

S7. References

(1) Firth, N. C.; Brown, N.; Blagg, J. Plane of Best Fit: A Novel Method to Characterize the Three-Dimensionality of Molecules. *J. Chem. Inf. Model.* 2012, 52, 2516–2525.

(2) O’Boyle, N. M.; Banck, M.; James, C. A.; Morley, C.; Vandermeersch, T.; Hutchison, G. R. Open Babel: An Open Chemical Toolbox. *J. Cheminformatics* 2011, 3, 33.

(3) Stewart, J. J. P. Optimization of Parameters for Semiempirical Methods VI: More Modifications to the NDDO Approximations and Re-Optimization of Parameters. *J. Mol. Model.* 2013, 19, 1–32.

(4) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Petersson, G. A.; Nakatsuji, H., et al. Gaussian 16, Revision C.01; Gaussian, Inc., Wallingford CT, 2016.

(5) Storn, R.; Price, K. Differential Evolution – A Simple and Efficient Heuristic for Global Optimization over Continuous Spaces. *J. Glob. Optim.* 1997, 11, 341–359.

(6) Virtanen, P.; Gommers, R.; Oliphant, T. E.; Haberland, M.; Reddy, T.; Cournapeau, D.; Burovski, E.; Peterson, P.; Weckesser, W.; Bright, J., et al. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nat. Methods* 2020, 17, 261–272.