UNIVERSIDADE FEDERAL DE SERGIPE
PRÓ-REITORIA DE PÓS-GRADUAÇÃO E PESQUISA PROGRAMA DE PÓS-GRADUAÇÃO EM QUÍMICA

EMANUEL FELIPE DOS SANTOS MATTOS

ESTUDO QSPR DA EFICIÊNCIA DE CÉLULAS SOLARES SENSIBILIZADAS POR CORANTES BASEADOS EM CARBAZOL

QSPR STUDY OF EFFICIENCY OF DYE-SENSITIZED SOLAR CELLS BY CARBAZOLE-BASED

UNIVERSIDADE FEDERAL DE SERGIPE

PRÓ-REITORIA DE PÓS-GRADUAÇÃO E PESQUISA PROGRAMA DE PÓS-GRADUAÇÃO EM QUÍMICA

EMANUEL FELIPE DOS SANTOS MATTOS

ESTUDO QSPR DA EFICIÊNCIA DE CÉLULAS SOLARES SENSIBILIZADAS POR CORANTES BASEADOS EM CARBAZOL

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Química, da Universidade Federal de Sergipe, para a obtenção do título de Mestre em Química

Orientador: Prof. Dr. Nivan Bezerra da Costa Junior

Coorientador: Prof. Dr. Carlos Raphael Araújo Daniel

QSPR STUDY OF EFFICIENCY OF DYE-SENSITIZED SOLAR CELLS BY CARBAZOLE-BASED

Master dissertation presented to the PostGraduate Program in Chemistry of the Federal University of Sergipe to obtain MSc in Chemistry

FICHA CATALOGRÁFICA ELABORADA PELA BIBLIOTECA CENTRAL
UNIVERSIDADE FEDERAL DE SERGIPE

Mattos, Emanuel Felipe dos Santos

M444e

Estudo QSPR da eficiência de células solares sensibilizadas por corantes baseados em carbazol / Emanuel Felipe dos Santos Mattos ; orientador Nivan Bezerra da Costa Junior - São Cristóvão, 2023.

110 f. : il.

Dissertação (mestrado em Química) – Universidade Federal de Sergipe, 2023.

1. Energia solar. 2. Células solares. 3. Corantes. Costa Junior, Nivan Bezerra orient. II. Título.

CDU 544.6

SERVIÇO PÚBLICO FEDERAL
MINISTÉRIO DA EDUCAÇÃO
UNIVERSIDADE FEDERAL DE SERGIPE
Programa de Pós-Graduação em Química
PPGQ

PPGQ

FOLHA DE APROVAÇÃO

Membros da Comissão Julgadora da Dissertação de Mestrado de Emanuel Felipe dos Santos Mattos apresentada ao Programa de Pós-Graduação em Química da Universidade Federal de Sergipe em 09/02/2023.

Prof. Dr. Nivan Bezerra da Costa Junior
Departamento de Química - UFS

Prof. Dr. Ricardo Oliveira Freire
Departamento de Química - UFS

Prof. Dr. Mario Ernesto Giroldo Valerio
Departamento de Física- UFS

AGRADECIMENTO

Aos meus familiares que me apoiaram nesse período.

A todos de contribuíram direta e indiretamente na minha formação pessoal e acadêmica.

Ao Professor Nivan pela orientação, conselhos e paciência durante a execução do trabalho.

Ao Professor Carlos pela coorientação.

Aos Professores Ricardo, Diogo e Mário pelas contribuições dadas durante a qualificação/defesa.

Aos Professores e amigos da pós que auxiliaram direta e indiretamente tanto no presente trabalho quanto na formação.

Ao sr. Helder e sra. Val sempre solícitos para solver as dúvidas sobre a parte administrativa.

À UFS pela estrutura, em especial o laboratório Pople.

À CAPES pelo auxílio financeiro.

RESUMO

Uma alternativa para os dispositivos fotovoltaicos comumente comercializado são as células solares sensibilizadas por corantes (DSSCs), porém elas ainda necessitam de melhorias para a obtenção de maiores eficiências de conversão. Nessas células, as otimizações são geralmente realizadas com a proposição de novos sensibilizadores. Neste sentido, no presente trabalho foi realizado um estudo QSPR da eficiência de DSSCs que utilizam corantes baseados em carbazol como sensibilizadores. Para isso, 2393 descritores moleculares foram obtidos para as 126 estruturas selecionas. Assim, por meio da regressão linear múltipla, quatro modelos foram obtidos, onde aquele com o maior poder de ajuste e predição ($R^2_{Treino} = 0{,}84$, $Q^2_{LOO} = 0{,}78$ e $R^2_{Pred} = 0{,}74$) foi interpretado. Dentre as características moleculares indicadas pelo modelo mais preditivo, estão apontadas a importância do comprimento da ponte π, do potencial de ionização e da presença de átomos eletronegativos. A partir dessas características, realizou-se um conjunto de modificações na estrutura do corante de maior PCE no conjunto utilizado, o que proporcionou identificar estruturas com PCE predito excedendo 11%.

Palavras-chave: Energia solar, DSSC, Corantes, Relação quantitativa.

ABSTRACT

An alternative to commonly marketed photovoltaic devices are dye-sensitized solar cells (DSSCs), but they still need improvements to provide greater conversion efficiencies. In these cells, optimizations are generally performed through proposition of new sensitizers. In this sense, in the present work a QSPR study was carried out on the efficiency of DSSCs that use carbazole-based dyes as sensitizers. For this, 2393 molecular descriptors were obtained from 126 selected structures. Thus, through multiple linear regression, four models were obtained, where the one with the highest adjustment and prediction power ($R^2_{Train} = 0.84$, $Q^2_{LOO} = 0.78$ e $R^2_{Pred} = 0.74$) was interpreted. Among the molecular characteristics indicated by the most predictive model, the importance of the π bridge length, the ionization potential and the presence of electronegative atoms are pointed out. Based on these characteristics, some modifications in dye's structure with the highest PCE in the dataset were made, which enabled the identification of structures with predicted PCE exceeding 11%.

Keywords: *Solar energy, DSSC, Dyes, Quantitative relationship.*

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA E ABREVIATURAS E SIGLAS

DSSC – Célula solar sensibilizada por corante (do inglês *"Dye-Sensitized Solar Cells)*

FTO – Óxido de Estanho dopado com Flúor (do inglês *"Fluorine doped Tin Oxide"*)

HCA – Análise do agrupamento hierárquico (do inglês *"Hierarchical Cluster Analysis"*)

MAE – Erro médio absoluto (do inglês *"Mean Absolute Error")*

MLR – Regressão linear múltipla (do inglês *"Multiple Linear Regression"*)

PCA – Análise do componente principal (do inglês *"Principal Components Analysis'*)

PCE – Poder de eficiência de conversão (do inglês *"Power Conversion Efficiency"*)

QSPR – Relação quantitativa estrutura-propriedade (do inglês *"Quantitative Structure Property Relationship"*)

RMSE – Raiz quadrada do erro médio (do inglês *"Root Mean Squared Error"*)

SUMÁRIO

1 INTRODUÇÃO

A demanda por energia cresce quase que continuamente ano a ano, tendo um aumento substancial a partir de meados do século passado. Sendo que grande parte da demanda é satisfeita com o uso de combustíveis fósseis, apesar do grande crescimento das fontes renováveis nas últimas décadas [1,2].

As energias descritas como renováveis são aquelas que apresentam um ciclo de renovação consideravelmente curto, sendo praticamente inesgotáveis, por exemplo, a eólica, hídrica e solar [1,3]. O sol fornece anualmente cerca de 10^{24} J de energia para a Terra [4], o que tem estimulado diversas pesquisas com o intuito de viabilizar o uso dessa enorme quantidade de energia. Como fruto dessas investigações, as células solares foram obtidas [5–7].

Uma célula solar é um dispositivo capaz de converter a energia dos raios solares incidentes em energia elétrica. A primeira célula que apresentou uma satisfatória eficiência de conversão (PCE, do inglês "*Power Conversion Efficiency"*), aproximadamente 6%, era baseada em silício e foi desenvolvida por Chapin e colaboradores em 1954 [8]. Desde então, o aprimoramento da composição e arquitetura resultaram em capacidades de conversão cada vez melhores, de modo que os dispositivos atualmente comercializados apresentam PCE entre 20-25% [6,9]. Além dessas, outras células têm chamado a atenção de pesquisadores nos últimos anos, tais como as células baseadas em perovskita [7,10], pontos quânticos [11] e corantes [4,5].

Em 1991, O'Regan e Grätzel obtiveram uma célula solar sensibilizada por corante (DSSC, do inglês "*Dye-Sensitized Solar Cell"*) com um PCE de 7% [12]. Por apresentarem um processo de produção simples e mais barato que o empregado para a obtenção daquelas baseadas em silício, as DSSCs vêm despertando o interesse dos pesquisadores [5,11]. Uma das principais vertentes de pesquisa é voltada para obtenção de novos corantes capazes de atuar em DSSC e proporcionar maiores eficiências de conversão.

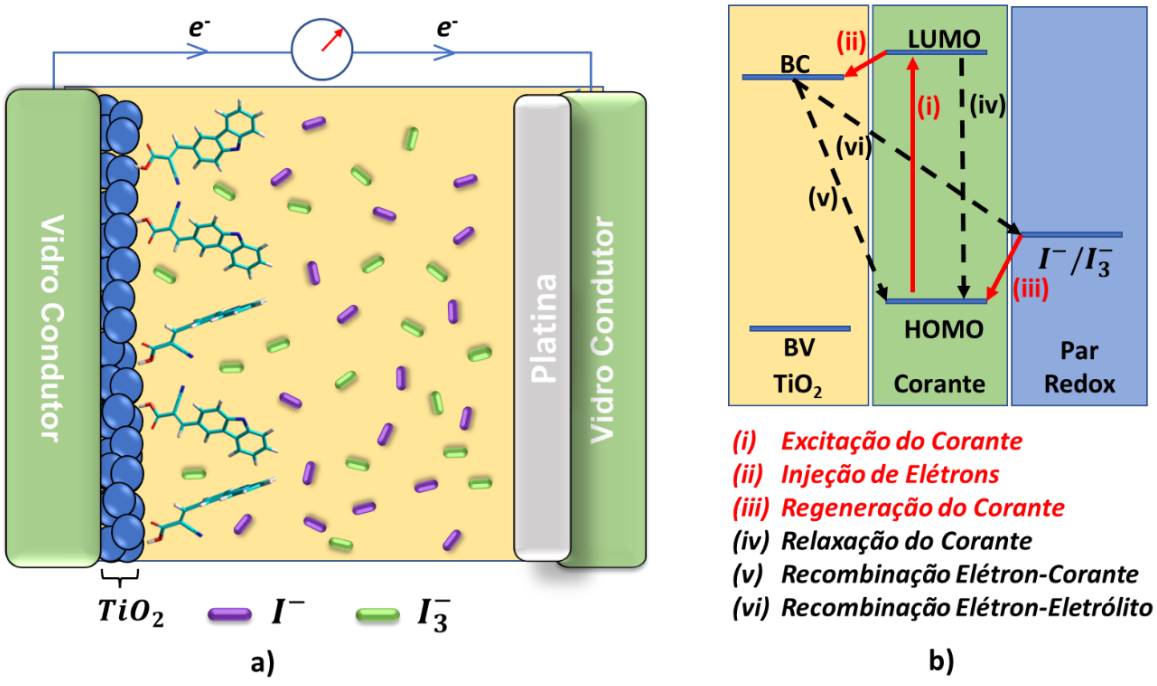
A modelagem QSPR (do inglês *"Quantitative Structure Property Relationship"*) tem se mostrado uma importante abordagem para a obtenção/modificação

de moléculas de maneira mais racional e econômica, além de possibilitar a predição da característica desejada para a estrutura proposta. Tal abordagem tem sido aplicada para a predição do PCE de DSSCs e para a proposição de novas estruturas de sensibilizadores [13–15].

1.1 Células Solares Sensibilizadas por Corantes

As DSSCs são constituídas por dois eletrodos (fotoanodo e contra eletrodo) e um eletrólito. Comumente, o fotoanodo é composto por um vidro de substrato condutor, o FTO (SnO_2:F), que contém uma camada de óxido semicondutor (normalmente o TiO_2) sensibilizado por um dado corante, enquanto o contra eletrodo é um vidro condutor contendo um filme de platina. O eletrólito normalmente adotado encontra-se no estado líquido, sendo composto por um solvente orgânico contendo como par redox o I^-/I_3^- [5,10,11], porém a literatura descreve a construção de células solares utilizando outros pares, como o Co^{2+}/Co^{3+} e Cu^{+}/Cu^{2+} [16,17]. A Figura 1-a apresenta a constituição de uma DSSC que adota o TiO_2 como óxido semicondutor e o I^-/I_3^- como par redox.

Figura 1 – Esquema (a) da constituição e (b) do funcionamento de uma DSSC.



As DSSCs geram corrente elétrica a partir do processo de transferência de

elétrons entre corante adsorvido, os eletrodos e par o redox [4,5]. O funcionamento da célula tem início na absorção da radiação na região do visível, hv_{vis} (400 nm ≤ λ ≤ 700 nm), que incide sobre o dispositivo, resultando na excitação do corante adsorvido (D_{ads}). Posteriormente, o corante excitado (D^*_{ads}) é oxidado, injetando elétrons na banda de condução (BC) do óxido semicondutor (e^-_{BC}). Tal como descrito nas equações 1 e 2, respectivamente.

$$D_{ads} + hv_{vis} \rightarrow D^*_{ads} \quad \text{(Eq. 1)}$$

$$D^*_{ads} \rightarrow D^+_{ads} + e^-_{BC} \quad \text{(Eq. 2)}$$

Para uma DSSC que utiliza I^-/I_3^- como par redox, o iodeto presente na solução atua como agente redutor e proporciona a regeneração do corante adsorvido (Eq. 3). Por fim, ocorre a redução do triiodeto no contra eletrodo, CE. (Eq. 4).

$$D^+_{ads} + \tfrac{3}{2} I^- \rightarrow D_{ads} + \tfrac{1}{2} I_3^- \quad \text{(Eq. 3)}$$

$$I_3^- + 2e^-_{CE} \rightarrow 3I^- \quad \text{(Eq. 4)}$$

As equações 1-4 descrevem o funcionamento ideal de uma DSSC, porém outras transferências de elétrons conduzem à redução da eficiência dessas células (Figura 1-b). Por exemplo, o corante excitado pode retornar ao seu estado fundamental e não injetar os elétrons na banda de condução do semicondutor. Ou ainda, os elétrons que foram injetados na banda de condução, ao invés de se difundirem sentido ao circuito externo, se recombinam com o corante oxidado ou com o eletrólito (Eq. 5 e 6).

$$D^+_{ads} + e^-_{BC} \rightarrow D_{ads} \quad \text{(Eq. 5)}$$

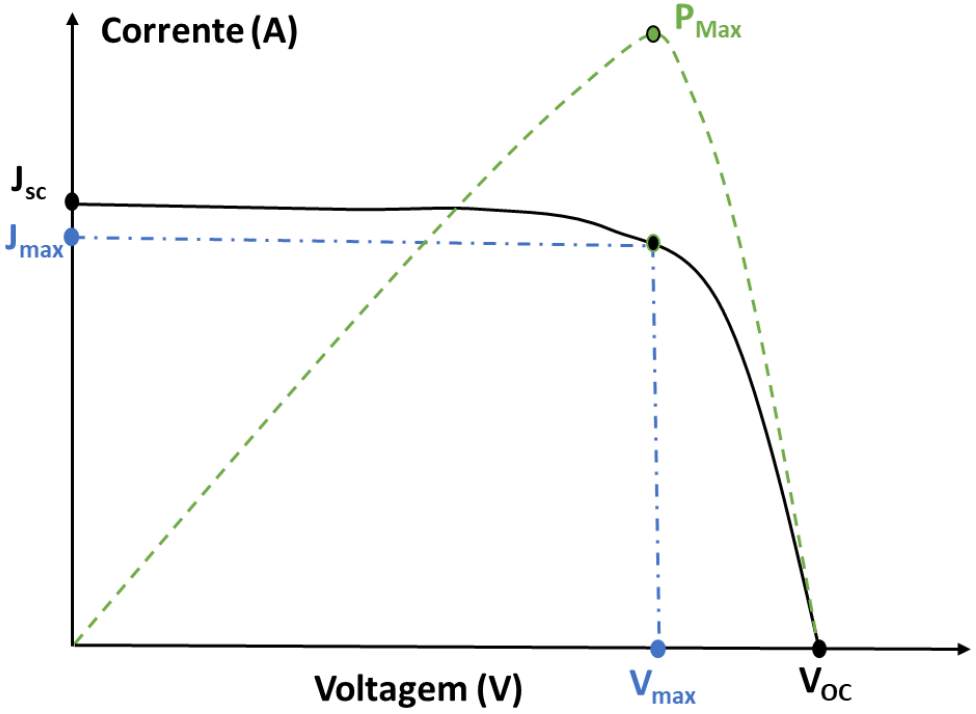
$$I_3^- + 2e^-_{BC} \rightarrow 3I^- \quad \text{(Eq. 6)}$$

Os processos de transição de elétrons descritos acima tendem a ocorrer quando há um ordenamento entre os orbitais de fronteira e as bandas do semicondutor (Figura 1-b). O padrão almejado é: (i) o orbital molecular não-ocupado de menor energia (LUMO) do corante combinando com a BC do semicondutor utilizado (possibilitando a injeção de elétrons) e (ii) orbital molecular ocupado de maior energia (HOMO) do corante com uma energia suficientemente baixa para permitir uma rápida regeneração do corante [18,19].

A conversão da energia solar em elétrica não ocorre de maneira ideal, ou seja, uma parcela da energia incidida sobre a célula não proporciona o fluxo de elétrons pelo circuito externo. Dessa maneira, é necessário mensurar a fração de energia solar que é efetivamente convertida em energia elétrica, então o PCE é calculado. Para obtê-lo, a potência máxima da célula (P_{max}) é determinada a partir da curva *corrente X tensão* (Figura 2), ou curva J-V, já que o PCE é definido como a razão entre a potência máxima da célula, P_{max}, e a potência da radiação incidida, P_{IN} [18,19].

Na curva J-V, a coordenada (V_{max},J_{max}) corresponde, respectivamente, aos valores de tensão e corrente que resultam na potência máxima da DSSC. Porém, outros dois pontos são importantes: (i) o J_{SC}, definido como a densidade de corrente de curto circuito (do inglês "*short-circuit current density*") quando o potencial é nulo; (ii) o V_{OC}, definido como a tensão de circuito aberto (do inglês "*open-circuit voltage"*).

Figura 2 – Curva J-V para performance de células fotovoltaicas.



Utilizando esses quatro parâmetros, é possível mensurar o chamado fator de preenchimento, FF (do inglês "*Fill Factor*"), o qual avalia o quão próximo do ideal (formato retangular) está uma dada curva J-V [20]. O FF pode ser calculado por:

$$FF = \frac{J_{\max} \times V_{\max}}{J_{SC} \times V_{OC}} \qquad \text{(Eq. 7)}$$

Desse modo, a eficiência de uma dada célula pode ser expressa em termos dos parâmetros J_{SC}, V_{OC} e FF, os quais são termos comumente apresentados na avaliação de um dispositivo fotovoltaico. Assim, o PCE pode ser calculado por:

$$PCE = \frac{P_{\max}}{P_{IN}} = \frac{J_{\max} \times V_{\max}}{P_{IN}} = \frac{FF \times J_{SC} \times V_{OC}}{P_{IN}} \qquad \text{(Eq. 8)}$$

1.1.1 Corantes

O corante é de fundamental importância para a conversão de energia solar em elétrica em uma DSSC. Por conseguinte, diversos estudos têm sido realizados com o propósito de compreender as características que tornam um corante um bom sensibilizador [21–25].

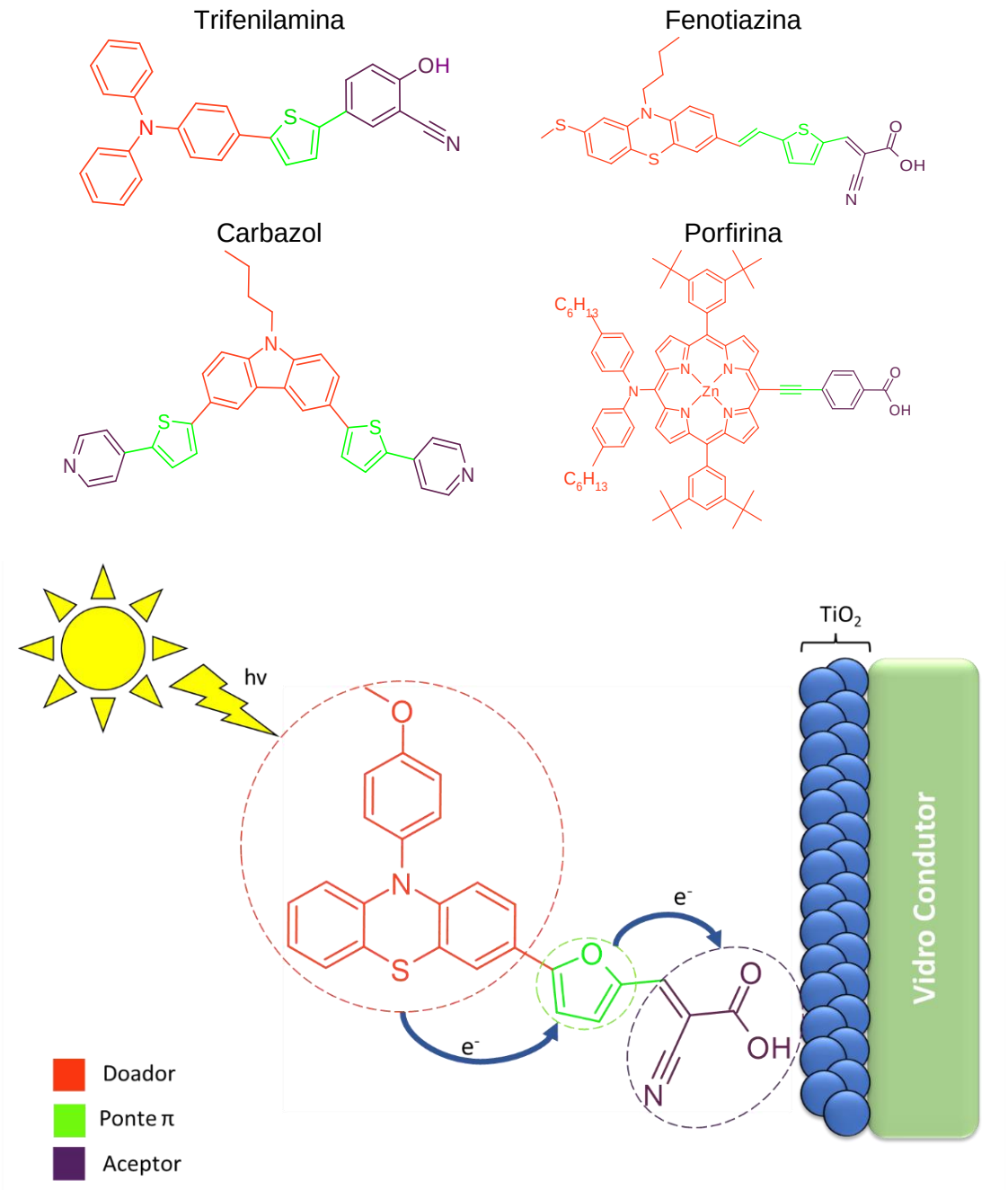
Desse modo, percebeu-se que um corante que atua como sensibilizador é composto de três grupos indispensáveis: o Doador (D); a Ponte/Espaçador π (π) e o Aceptor (A). A estrutura mais básica contendo esses componentes é conhecida como arquitetura D-π-A, porém outras estruturas, tais como D-A-π-A, (D-π-A)$_3$ e A-D-π-D-A têm sido estudadas mais recentemente [21–25].

Nesse contexto, várias espécies podem atuar como doadores ou espaçadores π, porém as estruturas contendo um sistema π conjugado são preferíveis. Por exemplo, a porfirina [26] e o carbazol [27] são comumente empregados como doadores, enquanto os anéis pirrólicos e tiofênicos [28] como espaçadores π.

Os aceptores apresentam os grupos denominados de ancoradores, tais como a carboxila, o fosfato e a piridina [29] que desempenham uma importante função no processo de adsorção do corante na superfície do semicondutor. A adsorção do corante pode ocorrer por alguns meios, como as interações de dispersão e ligações de hidrogênio, porém a formação de ligações covalentes envolvendo o aceptor (A) e a superfície do óxido é a via predominante [29]. Por isso, denomina-se a porção do sensibilizador que forma a ligação com a superfície do óxido (o aceptor) de grupo de ancoragem.

Na Figura 3 estão apresentadas algumas estruturas de sensibilizadores presentes na literatura [30–33] com a devida identificação dos grupos doador, ponte π e aceptor.

Figura 3 – Sensibilizadores encontrados na literatura e identificação dos grupos doador, ponte π e aceptor.



Quando o corante é excitado, ocorre uma transferência de carga intramolecular do doador para o aceptor e a posterior injeção dos elétrons na banda de condução do óxido [28,34]. Nesse sentido, estudos computacionais indicam que, em geral, a transferência de carga é dada pela transição HOMO→LUMO, com

outras contribuições menores, por exemplo, a HOMO→LUMO+1, HOMO-1→LUMO e HOMO-2→LUMO [5,27,28]. Assim, para que haja uma injeção de elétrons mais significativa na BC do semicondutor, proporcionada pela maior separação de carga, buscam-se sensibilizadores que apresentem o orbital HOMO com significativas contribuições do orbitais atômicos dos átomos do grupo doador e o LUMO com grandes contribuições de orbitais atômicos dos átomos do grupo aceptor [5].

Em suma, a literatura [5,19,35] indica que, para atuar em uma DSSC, o corante deve apresentar as seguintes características:

(i) capacidade de absorver significativamente em toda a região do visível;
(ii) possua no mínimo um grupo que permita sua ancoragem na superfície do semicondutor;
(iii) tenha o ordenamento dos seus orbitais HOMO e LUMO, em relação as bandas do semicondutor e do par redox, como apresentado na seção 1.1 (Figura 1-b).
(iv) o orbital HOMO sobre o grupo doador e o LUMO sobre o aceptor.

Utilizando-se desses princípios norteadores, foi possível projetar estruturas que resultaram em células com eficiência de conversão maiores que 9% [36–39], podendo alcançar valores entre 12-14% quando combinando com modificações nos eletrólitos e com o uso de coadsorventes [23,38].

Basicamente, as pesquisas que focam na estrutura do corante versam na utilização de dois tipos: aqueles que contém metais e os livres de metais. No primeiro tipo, destacam-se os compostos de coordenação de Ru(II) e alguns corantes porfirínicos. No segundo, ocorrem os compostos orgânicos sintéticos e naturais.

A síntese de corantes livre de metais para atuar como sensibilizadores em DSSCs tem crescido, pois são uma alternativa mais barata, quando comparada com aqueles contendo metais, e podem apresentar uma eficiência de conversão próxima dos 10% [19,23,39]. Além disso, os sensibilizadores orgânicos tendem a apresentar uma maior variabilidade estrutural que os corantes de complexos [5].

Comumente as modificações estruturais envolvendo os corantes orgânicos visam mudança nos grupos doadores [27,40], na ponte π [28,41] e nos grupos aceptores [42]. Por exemplo, Wang e colaboradores [43], ao trabalharem com corantes baseados em cumarina, realizaram a inserção de um segundo grupo CN a uma estrutura anteriormente sintetizada. Foi observado que a presença desse novo grupamento proporcionou o deslocamento da absorção máxima para o vermelho (de 511 para 552 nm), com um alto coeficiente de extinção molar ($9{,}74 \times 10^4$ $dm^3.mol^{-1}.cm^{-1}$). Além disso, o corante apresentou uma boa estabilidade de conversão, pois manteve um PCE constante de 6,2% durante 1000 horas de funcionamento.

Um outro estudo, realizado em 2018, referente à modificação de corantes baseados em cumarina de arquitetura D-π-A foi reportado por Dhar *et al.* [44]. Nesse estudo, os pesquisadores realizaram a síntese de uma série de sensibilizadores variando o número de anéis tiofênicos presentes na ponte π. Sendo observado que o aumento do número de unidades de tiofeno resultou no aumento do PCE, que saiu de 5,58% (1 anel) para 6,02% (3 anéis). Em termos de estabilidade, após 500 horas de uso, foi observada uma redução de apenas 0,2% na eficiência de conversão.

Em 2017, Ezhumalai *et al.* [45] realizou a síntese de um corante orgânico com uma das maiores eficiências de conversão (10,21%*)* sem o uso de coadsorventes. O sensibilizador sintetizado apresentava uma arquitetura D-π-A, na qual o grupo doador foi trifenilamina, o aceptor foi o ácido cianoacrílico e a ponte foi o tetratioaceno substituído com duas cadeias alquílicas ramificadas de oito carbono. Ademais, quando as cadeias ramificadas foram substituídas por cadeias lineares de quinze carbonos, o PCE reduziu para 9,02%. As inserções de cadeias alquílicas nas pontes tem como intuito reduzir as interações π-π entre os corantes, as quais quando presentes conduzem a diminuição do PCE [45].

Uma outra classe de corantes que tem sido aplicada como sensibilizadores em DSSCs são os baseados em carbazol, porém poucas células apresentam potencial de conversão superiores a 9%. Por exemplo, em 2007, Kim *et al*. [46] propuseram um conjunto de corantes N-aril substituídos, os quais se diferencia-

vam pela ponte e grupo doador. Em relação às modificações da ponte, os pesquisadores observaram que a inserção de um segundo anel tiofênico resultou no sutil aumento da eficiência de conversão (de 5,02% para 5,15%) quando o grupo doador é ácido cianoacrílico. Porém, para os corantes cujo os grupos doadores foram a rodanina, o aumento no número de anéis tiofênicos na ponte proporcionou a diminuição do PCE (de 1,75% para 0,55%).

Em 2013, Hu *et al.* [47] apresentou um outro conjunto de corantes baseados em carbazol, onde os autores buscaram aumentar a rigidez do grupo doador juntamente com a modificação das pontes π. Novamente, foi observado que o aumento no número de anéis tiofênicos resultou na elevação do PCE. Com essa abordagem foi possível a construção de células com uma eficiência de 7,15%. No mesmo ano, Liu e colaboradores [37] conseguiram obter um novo tipo de corante, também baseado em carbazol, de arquitetura D-π-A com uma cadeia cíclica envolvendo a ponte π. Tal sensibilizador apresentou uma satisfatória eficiência de conversão (9,20%) comparável aos corantes de trifenilamina e de rutênio [4,19].

Mais recentemente, Tian e colaboradores [27] realizaram a planarização do grupo doador, de um corante baseado em carbazol com arquitetura D-π-A. Os pesquisadores também observaram que o processo de planarização pouco afetou a quantidade de corante adsorvido na superfície do semicondutor, todavia proporcionou um leve acréscimo no J_{SC} e V_{OC} e, consequentemente, um aumento na eficiência de conversão de 6,16% para 7,21%. O aumento do PCE foi atribuído ao aumento da capacidade de absorção da luz na região do visível proporcionado pela planarização do grupo doador.

1.2 Relação Quantitativa Estrutura-Propriedade

A partir da fórmula estrutural de uma molécula é possível extrair um conjunto considerável de informações numéricas, por exemplo os átomos que a compõe, massa molecular, número ligações σ e π e, com a aplicação de métodos computacionais, a energia de formação, a energia dos orbitais de fronteira (HOMO e LUMO), cargas parciais dos átomos, dentre outras.

Ou seja, com o conhecimento da estrutura molecular e a aplicação de procedimentos matemáticos específicos ou experimentos padronizados, é possível extrair um conjunto de números úteis para descrever determinadas características moleculares [48]. Tais números são denominados de descritores moleculares, os quais podem ser classificados segundo a sua natureza, a citar os constitucionais, topológicos, eletrotopológicos, geométricos, solubilidade, químico-quânticos entre outros [48–51].

Existem milhares de descritores teóricos disponíveis na literatura que, no geral, foram desenvolvidos para aplicação em estudos de modelagem Relação (Quantitativa) Estrutura-Atividade, (Q)SAR, (do inglês, *"(Quantitative) Struture-Activity Relationships"*). Nesse tipo de modelagem, os descritores moleculares são utilizados para relacionar as características estruturais de uma dada molécula e uma atividade de interesse [48].

Quando a relação estabelecida é quantitativa, o resultado é um modelo estatístico devidamente validado que relaciona os descritores moleculares a uma atividade/propriedade molecular, ou seja, trata-se de expressão matemática onde o valor da atividade (variável dependente) é previsto a partir do conhecimento das variáveis independentes (descritores moleculares). Genericamente, o modelo QSAR é descrito pela eq. 9 [49].

$$P = f(D_1, D_2, D_3, \ldots D_n) \qquad \text{(Eq. 9)}$$

Em que P é a propriedade de interesse modelada em função dos D_i descritores moleculares, os quais são os pesados a partir do método estatístico empregado na construção do modelo. Esse tipo de abordagem permite, por exemplo a proposição de modificações estruturais para otimização da propriedade de interesse e a identificação de espécies que potencialmente não apresentam a propriedade desejada [49].

Com o intuito de estabelecer boas práticas voltadas para a construção de modelos QSAR, a OECD (do inglês *"Organization for Economic Co-operation and Development"*) propôs um conjunto de 5 princípios norteadores [52]: (i) atividade/Propriedade bem definida; (ii) algoritmo inequívoco; (iii) domínio de Apli-

cabilidade (DA) definido; (iv) apropriadas medidas de ajuste, robustez e preditividade; (iv) interpretação mecanística, quando possível.

O segundo princípio, algoritmo inequívoco, estabelece que se deve apresentar uma descrição transparente de todo o procedimento adotado para a obtenção do modelo. Isto é, o conjunto de estruturas utilizado, os valores de atividade/propriedade, como os descritores foram calculados, como foram gerados o conjunto de treino e teste, qual método matemático adotado para treinar o modelo etc [52].

As estruturas utilizadas na construção do modelo são um recorte de todo o espaço químico disponível. Por isso, recomenda-se definir o DA, pois ele representa o espaço químico estabelecido pelo conjunto de moléculas utilizado no treinamento do modelo. Ou seja, espécies que estão fora do DA são consideravelmente diferentes das estruturas utilizadas na obtenção do modelo, resultando em predições não confiáveis [52].

Uma das maneiras comumente empregadas para estabelecer o DA é utilizar o valor da alavancagem do indivíduo (ver seção 1.2.2), h_i. É esperado que os indivíduos apresentem uma alavancagem menor que a alavancagem limite $\boldsymbol{h^* = 3p/n}$, sendo $\boldsymbol{n}$ o número de objetos no conjunto de treinamento e $\boldsymbol{p}$ o número de descritores mais 1.

O quarto princípio descreve sobre a necessidade da adoção de métricas para o procedimento de validação interna (aplicada ao conjunto de treino) e externa (aplicada sobre o conjunto de teste) [52]. Para validação interna, exigem-se métricas de ajuste – capacidade do modelo de explicar a variabilidade da propriedade no conjunto de treino – e de robustez – estabilidade dos coeficientes de regressão do modelo após, por exemplo, a remoção de uma ou mais espécies do treinamento. E para a externa, necessita-se de parâmetros para avaliar a capacidade preditiva do modelo [52].

Quando se obtém um modelo QSAR devidamente validado, uma interpretação mecanística (princípio v) permite, quando possível, aprofundar o entendimento acerca da propriedade modelada, possibilitando a identificação/modelagem de novos compostos que possuam uma determinada característica [53–56].

As próximas três seções apresentam sucintamente alguns dos métodos utilizados em modelagem QSPR e que foram aplicados no presente trabalho.

1.2.1 PCA e HCA

Métodos de análise multivariada podem ser utilizados com o intuito de identificar padrões entre os objetos que compõem uma matriz com centenas/milhares de variáveis. Tais métodos, podem ser divididos em dois tipos: supervisionados e não supervisionados.

No primeiro, os objetos já apresentam uma classificação preestabelecida (bom/ruim, sim/não etc.), a qual é utilizada no estudo em questão. Nesse grupo, pode-se citar as técnicas do *k*-ésimo vizinho mais próximo, *k*-NN (do inglês "*Kth Nearest Neighbor"*) e análise do discriminante linear, LDA (do inglês *"Linear Discriminant Analysis"*) [57].

Já nos métodos não supervisionados, não há necessidade de os objetos possuírem uma classificação prévia, pois essa não é utilizada no processo de reconhecimento de padrões. Aqui, destacam-se as técnicas de análise do componente principal, PCA (do inglês *"Principal Component Analysis"*) e a análise do agrupamento hierárquico, HCA (do inglês *"Hierarchical Cluster Analysis"*) [57].

A PCA é um método de análise multivariada que permite a redução da dimensionalidade da matriz de dados e, consequentemente, uma interpretação mais facilitada dos dados com uma perda mínima de informação [58]. Ou seja, dada uma matriz com *n* objetos e *p* variáveis, $X_{n\,\mathrm{x}\,p}$, a PCA realiza a projeção desses *n* indivíduos em um subespaço de *k* dimensões (*k* < *p)*, chamadas de componentes principais, PCs.

Matematicamente, a PCA busca reescrever a matriz original $X_{n\,\mathrm{x}\,p}$ em duas outras: a matriz de escores, $T_{n\,\mathrm{x}\,k}$, e a matriz dos pesos, $L^{T}_{k\,\mathrm{x}\,p}$:

$$X_{n\,\mathrm{x}\,p} = T_{n\,\mathrm{x}\,k} \cdot L^{T}_{k\,\mathrm{x}\,\mathrm{p}} \qquad \text{(Eq. 10)}$$

Ao realizar a PCA, as componentes principais definem novos eixos não correlacionados, ou seja, a informação explicada por uma PC não está presente na outra. Tem-se também que, em termos de descrição da variabilidade dos dados da matriz original, as componentes seguem a seguinte ordem decrescente: PC1 > PC2 > PC3 > ... > PCk.

A matriz de escores, $T_{n\,\mathrm{x}\,k}$, é composta pelas coordenadas dos objetos nas PCs. Assim, pode-se identificar os possíveis padrões presentes na matriz original por meio da inspeção visual da distribuição dos objetos nos eixos formados pelas componentes principais [58]. E para justificar tal distribuição, observa-se o *plot* da matriz dos pesos, $L^{T}_{k\,\mathrm{x}\,p}$, pois ela, além de ser a responsável pela projeção dos objetos no subespaço *k*-dimensional, apresenta a contribuição de cada variável para formação das PCs [58].

A literatura apresenta alguns métodos destinados para a obtenção de ambas as matrizes, por exemplo a Decomposição por Valores Singulares, SVD (do inglês *"Singular Value Decomposition"*) e o algoritmo NIPALS (do inglês *"Non-linear iterative Partial Least-Squares"*) [59]. De maneira similar, diversos métodos para determinar o número de componentes principais que serão adotados na aproximação da matriz original estão presentes na literatura, por exemplo o critério de autovalores maiores que 1, teste *de elbow*, ou *scree,* e a porcentagem de variância explicada [59].

Embora a PCA normalmente permita identificar a formação de grupos de objetos em uma matriz de dados, ela não é um método voltado especificamente para o agrupamento. Para tal finalidade, utiliza-se métodos que visam justamente a formação de grupos baseado na similaridade/dissimilaridade entre os objetos.

O HCA é uma técnica de agrupamento hierárquico que opera por aglomeração dos objetos que compõem a matriz. Ou seja, para uma matriz de *N* objetos, existem inicialmente *N* grupos, porém os indivíduos vão sendo agrupados por meio de métricas de similaridade até que um grande grupo seja formado. A similaridade entre os objetos normalmente é obtida expressando a distância entre eles [57,59].

Alguns exemplos de distâncias normalmente utilizadas em HCA são a distância de Minkowski, a Manhattan e a euclidiana [59], as quais são calculadas segundo as equações 11 – 13 (Tabela 1), respectivamente. Dessas expressões, nota-se que a distância de Minkowski é uma generalização das distâncias Manhattan (p = 1) e euclidiana (p = 2). Assim, a distância entre os objetos *i* e *j*, d_{ij}, para um conjunto de *k* variáveis pode ser determinada.

Tabela 1 – Distâncias de Minkowski, Manhattan e Euclidiana.

Distância	Expressão	
Minkowski	$d_{ij} = \left[\sum_{k=1}^{K} \left\| x_{ik} - x_{jk} \right\|^{p} \right]^{1/p}$	(Eq. 11)
Manhattan	$d_{ij} = \sum_{k=1}^{K} \left\| x_{ik} - x_{jk} \right\|$	(Eq. 12)
Euclidiana	$d_{ij} = \left[\sum_{k=1}^{K} \left(x_{ik} - x_{jk} \right)^{2} \right]^{1/2}$	(Eq. 13)

As métricas mencionadas acima são utilizadas para descrever as distâncias entre os indivíduos que formam uma dada matriz, sendo a euclidiana a comumente adotada para essa finalidade. Já para a formação dos grupos, isto é, para medir a distância entre os agrupamentos, outros métodos são utilizados. Os principais são [57,59]:

(i) **Método simples:** Para esse método, a distância entre dois grupos quaisquer é dada pela a menor distância entre os indivíduos dos dois grupos;

(ii) **Método completo:** É o oposto do método simples, ou seja, a distância entre os grupos é definida pela maior distância entre os indivíduos dos dois grupos;

(iii) **Método da média:** A distância entre os grupos é obtida por meio da média das distâncias entre todos os pares de indivíduos que compõem os grupos.

(iv) **Método *ward*:** Nesse método, as distâncias entre os grupos são calculadas a partir das somas quadráticas entre os centroides de cada grupo.

Para os agrupamentos hierárquicos, o método *ward* normalmente resultam dendogramas bem estruturados com uma tendência em formar grupos de tamanhos aproximadamente similares [59,60].

1.2.2 Regressão Linear Múltipla

Em alguns momentos, utilizar apenas um descritor para modelar uma determinada propriedade não é possível, ou seja, são necessários dois ou mais DMs para que a relação estrutura-atividade seja realmente estabelecida. Por esse motivo, técnicas de regressão multivariada são comumente adotadas em modelagem QSPR, a citar o MLR, Regressão Linear Múltipla (do inglês *'Multiple Linear Regression")*.

O MLR é uma importante técnica de regressão no campo de modelagem QSPR, pois possibilita, quando possível, uma interpretação mecanística mais direta do modelo [61]. Esse fato está intrinsecamente associado com a forma final do modelo, a qual é dada por:

$$\hat{Y} = \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + \alpha_3 X_3 \cdots + \alpha_n X_n \qquad \text{(Eq. 14)}$$

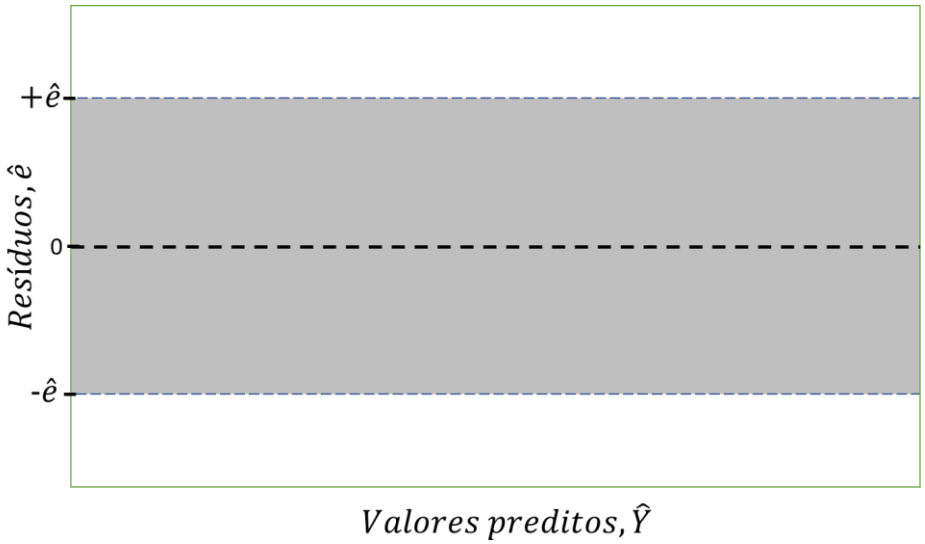
Sendo a variável dependente, $\hat{Y}$, o valor predito da propriedade Y, as variáveis independentes, $X_1,\ X_2, \ldots, X_n$, os descritores moleculares que compõem o modelo e $\alpha_0, \alpha_1, \ldots, \alpha_n$ são os parâmetros de regressão. Matematicamente, a regressão linear busca um conjunto de parâmetros $\{\alpha_0, \alpha_1, \ldots, \alpha_n\}$, para uma dada coleção de descritores $\{X_1,\ X_2, \ldots, X_n\}$, que minimize a soma quadrática dos resíduos, $\sum\left(Y_i - \hat{Y}_i\right)^2$ [57,62]. Ou seja, o MLR opera com o intuito de obter os mínimos quadrados dos resíduos, isto é:

$$\frac{\partial}{\partial \alpha_j}\left[\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2\right] = 0 \qquad \text{(Eq. 15)}$$

Porém, para que a regressão seja performada, assume-se que as variáveis independentes não possuem erros e que os resíduos são aleatória e identicamente distribuídos em torno de zero e apresentam uma variância constante e igual a σ_y^2 (são homocedásticos) [57,62]. Assim, observar o gráfico dos resíduos em função dos valores preditos, $\hat{Y}$, permite a extração de algumas conclusões acerca do modelo obtido, por exemplo se os resíduos apresentam uma variância não constante, se há erro na estimativa da constante α_0, a necessidade de modificação da ordem do modelo ou inserção de mais variáveis [57,62].

Nesse sentido, espera-se que o gráfico de resíduos *vs.* valores preditos apresente um perfil característico quando o modelo for razoavelmente adequando (Figura 4). Ou seja, os resíduos devem estar distribuídos, ao redor do zero, em uma região retangular compreendida entre ±*ê* e qualquer modificação pronunciada nesse padrão deve ser analisada [62].

Figura 4 – Perfil característico do gráfico resíduos vs. valores preditos.



O grande desafio na obtenção de modelos QSPR via MLR e a seleção de uma coleção adequada de descritores moleculares $\{X_1,\ X_2, \dots, X_n\}$. Assim, a seleção de variáveis é uma etapa crucial, e a mais desafiadora, na obtenção dos modelos preditivos. A literatura apresenta um conjunto considerável de métodos para selecionar variáveis, tais como o *stepwise* [63] e o algoritmo genético [64].

1.2.3 Validação de modelos

O processo de validação interna e externa é uma etapa fundamental na construção de modelos QSPR, pois, ao satisfazer um conjunto de métricas, é possível dizer que o modelo realiza predições suficientemente confiáveis. Para uma validação adequada, tanto o conjunto de treinamento quanto o teste devem ser avaliados por meio de um conjunto de métricas [52].

Para validar internamente um modelo, alguns dos parâmetros/métodos que podem ser utilizados são os coeficientes de determinação (R^2_{Treino}, R^2_{Adj} e Q^2), medidas de erro (RMSE e MAE), métricas de r^2_m e *y-randomization*. Assim, para um modelo de *k* parâmetros obtido a partir de um conjunto de treinamento com

N objetos, algumas das métricas para validação interna são calculadas segundo as expressões apresentadas na Tabela 2. A observação conjunta dessas métricas permite analisar a capacidade de ajuste do modelo analisado.

Tabela 2 – Metricas para validação interna.

Métrica	Expressão	
R^2_{Treino}	$R^2_{Treino} = 1 - \frac{\sum_{i=1}^{N}\left(Y_i - \hat{Y}_i\right)^2}{\sum_{i=1}^{N}\left(Y_i - \bar{Y}_i\right)^2}$	(Eq. 16)
R^2_{Adj}	$R^2_{Adj} = 1 - \frac{N-1}{N-k-1}(1 - R^2_{Treino})$	(Eq. 17)
Q^2	$Q^2 = 1 - \frac{\sum_{i=1}^{N}\left(Y_i - \hat{Y}_{i(Modelo\ cruzado)}\right)^2}{\sum_{i=1}^{N}\left(Y_i - \bar{Y}_i\right)^2}$	(Eq. 18)
RMSE	$RMSE = \sqrt{\frac{\sum_{i=1}^{N}\left(Y_i - \hat{Y}_i\right)^2}{N}}$	(Eq. 19)
MAE	$MAE = \frac{\sum_{i=1}^{N}\left\|Y_i - \hat{Y}_i\right\|}{N}$	(Eq. 20)
VIF	$VIF_i = \frac{1}{1 - R_i^2}$	(Eq. 21)

Os coeficientes de determinação R^2_{Treino} e R^2_{Adj} servem como um parâmetro indicador do quão bem os valores das propriedades dos *N* objetos são reproduzidos pelo modelo [65], a qual é reforçada por valores de RMSE e MAE suficientemente baixos. O R^2_{Treino} e R^2_{Adj} se diferenciam pelo caráter punitivo que o R^2_{Adj} apresenta em relação ao aumento da complexidade do modelo, assim ele permite avaliar modelos com diferentes números de variáveis. Espera-se ainda que $R^2_{Treino} - R^2_{Adj} < 0{,}3$, pois é um indicativo de que o número de parâmetros no dado modelo é adequado [66].

Dessas métricas, o Q^2 merece uma atenção especial, pois trata-se do coeficiente de determinação da validação cruzada. Na validação cruzada, o conjunto de treinamento é dividido em *p* grupos, dos quais *p - 1* são utilizados para recalcular os parâmetros de regressão e o grupo restante é utilizado para avaliar a

capacidade preditiva do novo modelo gerado, esse procedimento é repetido até que todos os *p* grupos tenham sido testados [65]. Quando *p* = *N*, a validação cruzada é denominada de LOO (do inglês *"leave-one-out"*), porém quando *p* ≠ *N* e cada grupo p_i contém *n* indivíduos, ela é nomeada de LNO (do inglês *"leave-n-out"*) [65].

O Q^2 é normalmente relatado como uma métrica capaz de indicar a capacidade preditiva de um modelo: (i) quando Q^2 é pequeno, tem-se o indicativo de que o modelo não apresenta um poder de previsão adequado; (ii) porém quando Q^2 é grande, possivelmente o modelo tem uma boa capacidade de previsão [66,67]. Provavelmente, um modelo é composto por descritores relevantes e apresenta uma potencial capacidade preditiva se $Q^2 > 0{,}5$ e a diferença entre R^2_{Treino} e Q^2 seja menor que 0,2 – 0,3 [65].

Roy e colaboradores [68,69] inicialmente propuseram a métrica r_m^2 para validações externas, porém com uma modificação na metodologia, ela pôde ser empregada para validar internamente um modelo. No contexto da validação interna, tal métrica se apresenta como $\overline{r^2_{m\,(LOO)}}$ e $\Delta r^2_{m\,(LOO)}$, as quais são obtidas a partir dos coeficientes de determinação para as regressões entre os valores observados, Y, e preditos pela validação cruzada LOO, $\hat{Y}_{(LOO)}$, para o conjunto de treino e, segundo elas, um modelo pode ser considerado internamente validado se $\overline{r^2_{m\,(LOO)}} > 0{,}5$ e $\Delta r^2_{m\,(LOO)} < 0{,}2$. Roy *et al.* [70] ainda mostraram que utilizar os valores Y e $\hat{Y}_{(LOO)}$ escalados pela amplitude resultam em $\overline{r^2_{m\,(LOO)}}$ e $\Delta r^2_{m\,(LOO)}$ mais punitivos em relação às métricas obtidas com os valores não escalados.

Na tabela 3 estão apresentados os termos necessários para calcular as métricas $\overline{r^2_{m\,(LOO)}}$ e $\Delta r^2_{m\,(LOO)}$. Desses termos, $r^2_{(LOO)}$, $R^2_{o(LOO)}$ e $R'^2_{o(LOO)}$ são os coeficientes de determinação da referida regressão, sendo que $r^2_{(LOO)}$ refere-se a regressão com intercepto, enquanto os outros são obtidos sem intercepto – métodos de regressão linear que não adotam o intercepto são denominados de regressão através da origem. Vale destacar que a diferença entre $R^2_{o(LOO)}$ e $R'^2_{o(LOO)}$ repousa sobre o fato de que $R^2_{o(LOO)}$ está associado à regressão em um gráfico do tipo $Y\ x\ \hat{Y}_{(LOO)}$, enquanto o $R'^2_{o(LOO)}$ é para o gráfico $\hat{Y}_{(LOO)}\ x\ Y$ [67].

Tabela 3 – Equações para obtenção das métricas r_m^2 para validação interna.

Métrica	Expressão	
$r^2_{(LOO)}$	$r^2_{(LOO)} = \dfrac{\left[\sum\left(Y_i - \overline{Y}_i\right)\left(\hat{Y}_{i(LOO)} - \overline{\hat{Y}}_{(LOO)}\right)\right]^2}{\sum\left(Y_i - \overline{Y}_i\right)^2 \sum\left(\hat{Y}_{i(LOO)} - \overline{\hat{Y}}_{(LOO)}\right)^2}$	(Eq. 21)
$R^2_{o(LOO)}$	$R^2_{o(LOO)} = 1 - \dfrac{\sum\left(Y_i - k\hat{Y}_{i(LOO)}\right)^2}{\sum\left(Y_i - \overline{Y}_i\right)^2}, \quad k = \dfrac{\sum\left(Y_i - \hat{Y}_{i(LOO)}\right)}{\sum \hat{Y}^2_{i(LOO)}}$	(Eq. 22)
$R'^2_{o(LOO)}$	$R'^2_{o(LOO)} = 1 - \dfrac{\sum\left(\hat{Y}_{i(LOO)} - k'Y_i\right)^2}{\sum\left(\hat{Y}_{i(LOO)} - \overline{\hat{Y}}_{i(LOO)}\right)^2}, \quad k' = \dfrac{\sum\left(Y_i - \hat{Y}_{i(LOO)}\right)}{\sum Y_i^2}$	(Eq. 23)
$r^2_{m(LOO)}$	$r^2_{m(LOO)} = r^2_{(LOO)}\left(1 - \sqrt{r^2_{(LOO)} - R^2_{o(LOO)}}\right)$	(Eq. 24)
$r'^2_{m(LOO)}$	$r^2_{m(LOO)} = r^2_{(LOO)}\left(1 - \sqrt{r^2_{(LOO)} - R'^2_{o(LOO)}}\right)$	(Eq. 25)
$\overline{r^2_{m\,(LOO)}}$	$\overline{r^2_m} = \dfrac{r^2_{m(LOO)} + r'^2_{m(LOO)}}{2}$	(Eq. 26)
$\Delta r^2_{m\,(LOO)}$	$\Delta r^2_m = \left\lvert r^2_{m(LOO)} - r'^2_{m(LOO)} \right\rvert$	(Eq. 27)

Com um intuito de avaliar a significância do modelo obtido, o teste de *y-randomization* normalmente é aplicado. Nesse método, o vetor de propriedade, o vetor y, é permutado um determinado número de vezes enquanto as posições dos descritores ficam inalteradas, e em cada permutação um modelo paralelo é obtido. Espera-se que os modelos obtidos com o vetor randomizado apresente os coeficientes de determinação (R^2_{Treino} e Q^2) menores que o do modelo 'real' [65].

Existe um conjunto considerável de métricas que podem ser adotadas para a validação externa, dentre elas a mais presente é o coeficiente de determinação para a predição, R^2_{Pred}, a qual é uma expressão bastante similar a Eq. 16, porém aplicada para o grupo de teste:

$$R^2_{Pred} = 1 - \frac{\sum_{i=1}^{Teste}\left(Y_i - \hat{Y}_i\right)}{\sum_{i=1}^{Teste}\left(Y_i - \overline{Y}_{Treino}\right)} \qquad \text{(Eq. 27)}$$

Sendo Y_i e $\hat{Y}_i$ o valor observado e predito, respectivamente e $\overline{Y}_{Treino}$ é a

média do valor observado do conjunto de treinamento. Segundo tal métrica, um modelo pode ser descrito como preditivo se R^2_{Pred} ≥ 0,5 [71]. Outras métricas, nomeadas de Q^2_{F2} e Q^2_{F3}, baseadas em pequenas modificações da Eq. 27, também costumam ser adotadas para avaliação da capacidade preditiva de um modelo [53].

Os métodos de regressão através da origem também costumam ser adotados para validação externa, por exemplo Golbraikh e Tropsha [67] propõem o uso dessa técnica para avaliar a previsibilidade do modelo. Segundo esses autores, ao realizar tal regressão em um gráfico do tipo $Y_{teste}\ x\ \hat{Y}_{teste}$, deve haver uma diferença menor que 10% entre r^2 e R^2_o. Além disso, o coeficiente angular para a regressão, *k*, deve estar compreendido no intervalo [0,85, 1,15].

Em adição, Roy *et al.* [69] propuseram o r^2_m aplicado para validação externa. Essas métricas podem ser calculadas segundo as equações 21, 22 e 24, porém os valores preditos não são obtidos pela validação cruzada, mas pelo modelo proposto. Além do mais, para as validações externas utilizando a métrica r^2_m, há necessidade de escalar os vetores na amplitude [69].

Nesse sentido, empregado os métodos e métricas apresentadas acima, o presente estudo consiste em obter modelos QSPR robustos e preditivos voltados para a eficiências de DSSCs que utilizam corantes orgânicos baseados em carbazol, pois é uma das classes de sensibilizadores que necessita de melhorias e tem sido pouco explorada por essa abordagem.

2 OBJETIVOS

2.1 Objetivo Geral

Realizar um estudo QSPR das performances fotovoltaicas de células solares sensibilizadas por corantes orgânicos baseados em carbazol.

2.2 Objetivos Específicos

- Selecionar estruturas de corantes baseados em carbazol utilizados na sensibilização de DSSCs construídas e caracterizadas sob condições similares;
- Calcular os descritores estruturais (2D e 3D), eletrônicos e espectrais para o conjunto de estruturas dos corantes orgânicos selecionados;
- Aplicar métodos de seleção de variáveis que possibilitem a obtenção de modelos robustos e preditivos para a eficiência fotovoltaica de uma DSSC sensibilizada por corantes baseados em carbazol;
- Validar interna e externamente o(s) modelo(s) obtido(s) utilizando métricas e métodos presentes na literatura;
- Buscar compreender a relação entre a estrutura do corante e a eficiência de conversão de uma DSSC que o utilize;

3 METODOLOGIA

3.1 Obtenção das estruturas dos corantes e performances fotovoltaicas

As estruturas e os dados experimentais (PCE, J_{sc}, V_{OC} e FF) utilizados no presente trabalho foram obtidos do banco de dados *Open Acess* disponibilizado por Venkatraman e Chellapan [72]. Além de possuir o carbazol como grupo doador, outros critérios para a seleção dos corantes foram utilizados:

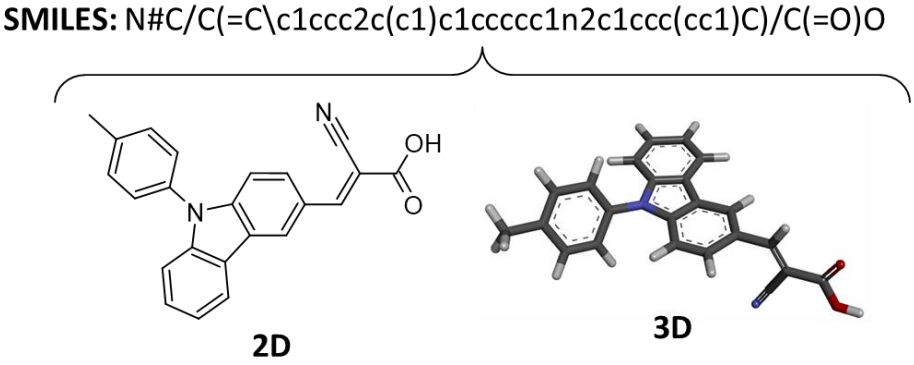
(i) TiO_2 como semicondutor no fotoanodo;
(ii) Filme de platina no contra eletrodo;
(iii) O par I^-/I_3^- como eletrólito;
(iv) Ácido cianoacrílico como grupo de ancoragem.

O banco disponibilizava 521 corantes baseados em carbazol, porém após a aplicação dos critérios mencionados acima, 126 foram selecionados (Apêndice 8.1).

3.2 Otimização Estrutural e cálculo dos descritores moleculares

Após a seleção, as estruturas bi e tridimensionais dos corantes foram obtidas a partir dos SMILES (do inglês *"simplified molecular-input line-entry system"*) disponibilizados no banco de dados [72], adotando o ChemSketch versão 2021.1.0 para conversão. A Figura 5 apresenta o SMILES de uma estrutura de um corante e as respectivas estruturas 2D e 3D.

Figura 5 – Estruturas 2D e 3D de um corante baseado em carbazol e seu SMILES.



Para a otimização da geometria molecular, o pacote computacional MOPAC2016 [73] foi empregado. Todas 126 estruturas foram otimizadas em fase

gasosa adotando o modelo semiempírico RM1, GNORM = 0,1 como critério de convergência e, para calcular a polarizabilidade, a palavra-chave POLAR foi utilizada.

Sabendo que uma forte absorção na região do visível é uma importante característica que o corante deve possuir para atuar como sensibilizador, buscou-se incluir esses dados como descritores moleculares. Porém, como as tabelas com as absorções experimentais não estavam disponíveis, optou-se por trabalhar com os espectros de absorção teóricos considerando o mesmo solvente utilizado na caracterização do corante no artigo de origem. Para considerar o efeito do solvente, as palavras-chave EPS e N**2 (constante dielétrica e índice de refração, respectivamente) foram empregas. A Tabela 4 apresenta as constantes dielétricas e os índices de refração utilizados nos cálculos.

Tabela 4 – Constantes dielétricas e índices de refração para cada solvente considerado.

Solvente	Constante Dielétrica	Índice de Refração
Acetronitrila	37,5	1,34
Diclorometano	8,93	1,42
Etanol	24,5	1,36
Metanol	32,7	1,32
THF	7,58	1,40
Triclorometano	4,81	1,40

Após a otimização das geometrias dos corantes, os descritores estruturais (2D e 3D) foram calculados empregando o *software* PaDEL – Descriptor [74]. Além desses descritores, a área e volume acessível ao solvente, potencial de ionização, eletroafinidade, eletronegatividade, dureza, eletropositividade, momento dipolo (componentes e total), o alfa médio isotrópico e as cargas parciais dos átomos dos grupos carboxila e cianeto foram obtidos a partir do MOPAC2016.

Para obter os espectros teóricos, os estados excitados das estruturas otimizadas foram calculados adotando o hamiltoniano INDO/S com abordagem de interação simples implementado no MOPAC2016. Relembrando que o método COSMO foi utilizado para considerar o efeito do solvente. Então os espectros teóricos foram obtidos através de ajustes Lorentzianos com largura de banda à meia altura de 25 cm^{-1}.

3.3 Divisão dos grupos de treinamento e teste

Como intuito de obter grupos de treino e teste que representassem de maneira mais adequada o conjunto de moléculas utilizadas para a modelagem, uma análise de agrupamento hierárquico foi realizada sobre uma análise de componentes principais. Ambas foram performadas utilizando a linguagem R no *RStudio* [75,76] adotando o pacote *FactoMineR* [77].

Para a PCA, as variáveis experimentais PCE, J_{SC}, V_{OC} e FF foram consideradas variáveis quantitativas suplementares. Então, um conjunto de componentes principais que explicasse no mínimo 90% da informação da matriz original foi selecionada e a HCA foi performada sobre essas componentes. As semelhanças foram calculadas a partir das distâncias euclidianas, empregando o método *ward* para criação dos *cluster*s.

Após o agrupamento, uma amostragem aleatória foi realizada em cada um dos grupos formados, de modo que 3/4 de cada *cluster* fosse destinado à formação do grupo de treino (os corantes com os valores máximo e mínimo de PCE foram forçados a compor esse grupo) e o restante para o grupo de teste.

3.4 Desenvolvimento e validação dos modelos

Antes de realizar a seleção de variáveis, a matriz de dados original foi reduzida utilizando os critérios de baixa variância ($\sigma^2 < 10^{-6}$) e correlação ($r > 0,95$). Na eliminação por correlação do par de variáveis com $r > 0,95$, aquela mais correlacionada com a propriedade a ser modelada foi mantida na matriz de dados.

Assim como o PCA e HCA, a obtenção dos modelos foi realizada no RStudio [75,76]. O método de ajuste adotado para a modelagem foi a regressão linear múltipla e as abordagens de seleção de variáveis foram o algoritmo genético e o *best subsets* disponíveis nos pacotes *gaselect* [78] e *leaps* [79], respectivamente. Outro importante pacote utilizado foi o *metrics* [80], empregado para o cálculo do RMSE, MAE e VIF.

Primeiramente, buscou-se definir o conjunto de parâmetros para o algoritmo genético modelar o PCE como variável resposta. Para isso, foram modificados o

tamanho da população (50 a 250, em intervalos de 100), número de gerações (50 a 1000, em intervalos de 50), a probabilidade de mutação (0,01, 0,05, 0,1 e 0,15) e o método de *crossover* (*single* e *random*), deixando o número de variáveis fixados entre 5 e 12. Cada configuração foi repetida 50 vezes, sendo que o melhor R^2 e a média dos R^2 de cada iteração foram registrados. Assim, o GA foi configurado com base na média dos melhores R^2 e melhores médias de R^2 médio.

Um *script* (Apêndice 8.5) foi escrito para automatizar o processo de obtenção dos modelos. Em um primeiro momento, o *script* utilizou o algoritmo genético (10000 iterações) para selecionar apenas os modelos que obedecessem às métricas: $R^2_{Treino} \geq 0,6$, $R^2_{Pred} \geq 0,5$ e, para cada descritor, valor-p ≤ 0,05 e VIF ≤ 5. Em seguida, os modelos obtidos via GA foram refinados adotando o método *best subsets*, com 2000 iterações. Sendo considerado como aprimorado os modelos que tiveram $R^2_{Treino} \geq 0,7$, $R^2_{Pred} \geq 0,65$, valor-p ≤ 0,05 e VIF ≤ 5, para cada descritor e no máximo 14 descritores.

Assim, os modelos obtidos foram internamente validados por meio das métricas R^2_{Treino}, Q^2_{LOO}, métricas r^2_m e *Y-randomization* (500 iterações). Já a validação externa foi realizada através do R^2_{Pred} e métricas baseadas em regressão através da origem para os valores observados e preditos [67,69].

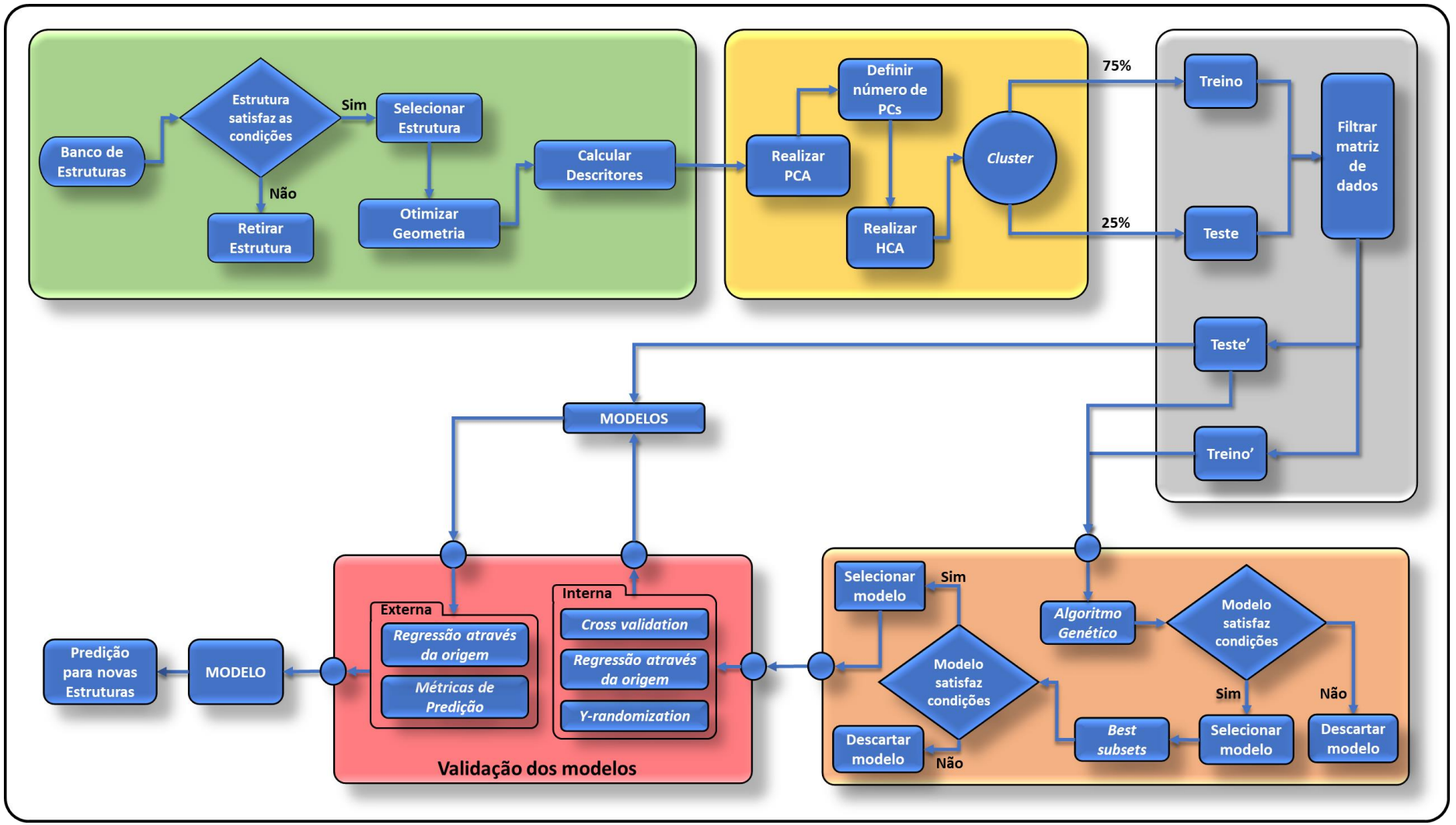
3.5 Modificações estruturais

Do conjunto de 126 corantes selecionados, aquele presente na DSSC de maior eficiência foi selecionado para um conjunto de modificações estruturais visando o aumento do PCE da célula.

As regiões e os grupos escolhidos para realizar as modificações foram baseados nas características estruturais indicadas pelo modelo de maior poder de ajuste e predição.

O fluxograma apresentado na Figura 6 apresenta resumidamente a metodologia empregada no trabalho.

Figura 6 – Fluxograma apresentando as etapas empregadas no trabalho.



4 RESULTADOS E DISCUSSÃO

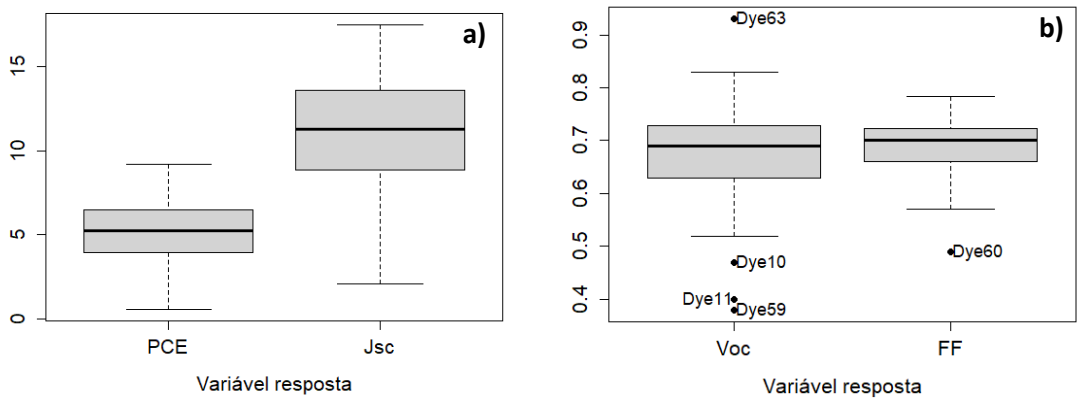
4.1 Matriz de dados

4.1.1 Variáveis Resposta

Como mencionado na seção 3.1, a aplicação dos critérios para a seleção dos corantes no banco de dados resultou na escolha de 126 corantes baseados em carbazol. Assim, a Figura 7 apresenta os gráficos *boxplots* das variáveis experimentais desses corantes, nela é possível visualizar que tanto o PCE quanto o J_{sc} (Fig. 7-a) apresentam uma distribuição aproximadamente simétrica em relação à mediana, com os valores compreendidos entre 0,57% ≤ PCE ≤ 9,2% e 2,06 mA.cm^{-2} ≤ J_{SC} ≤ 17.49 mA.cm^{-2} e sem a presença de *outliers*.

Por outro lado, o V_{OC} e FF (Fig. 7-b), que também estão distribuídos de maneira aproximadamente simétrica, apresentam alguns valores que diferem significativamente do conjunto.

Figura 7 – *Bloxplot* para as variáveis resposta da matriz de dados.



Apesar desses corantes (*Dye 10, Dye 11, Dye59, Dye 60 e Dye 63*) apresentarem V_{OC} ou FF que destoam do conjunto selecionado, eles foram mantidos na matriz de dados. Desse modo, a tensão de circuito aberto e fator de preenchimento estão distribuídos no intervalo 0,38 V ≤ V_{OC} ≤ 0,93 V e 0,49 ≤ FF ≤ 0,78, respectivamente.

4.1.2 Variáveis Preditoras

Como já mencionado, após a otimização estrutural dos corantes selecionados, os cálculos dos descritores 2D e 3D foram realizados através do PaDEL-Descriptor [74]. Esse *software* é capaz de calcular um total de 1875 descritores (1444 2D e 431 3D).

Sabendo que uma intensa absorção na região do visível é uma característica que o sensibilizador deve possuir para atuar em uma DSSC [5], o espectro de absorção teórico de cada corante selecionado foi calculado (Apêndice 8.2). Com a inspeção visual desses espectros, verifica-se que as absorções representativas estão dispostas entre 400 – 600 nm. Dessa maneira, apenas as absorções compreendidas nessa faixa (500 comprimentos de onda) foram utilizadas como descritores moleculares neste estudo.

Com isso, a matriz de dados apresentou uma dimensão de 126 x 2397, 126 objetos e 2397 variáveis, sendo as quatro respostas (PCE, J_{SC}, V_{OC} e FF) e 2393 descritores (1875 do PaDEL,18 do MOPAC2016 e intensidade para 500 comprimentos de onda).

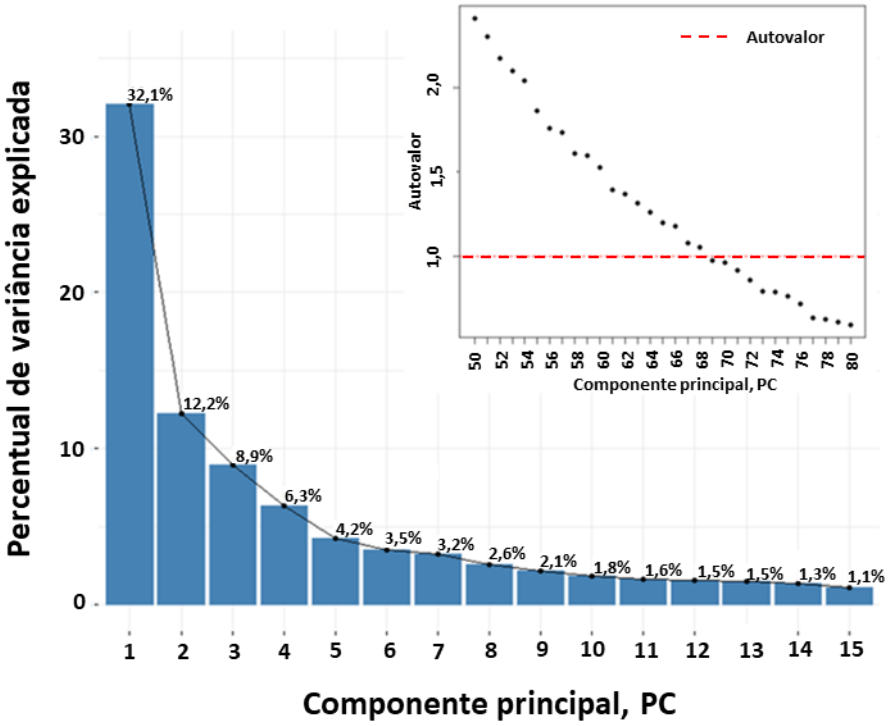
4.2 HCA e a obtenção dos grupos de Treino e Teste

A análise de componentes principais é um método de análise multivariada que permite, além da diminuição do ruído do conjunto de dados, a extração das informações mais relevantes e a simplificação na descrição da matriz e o reconhecimento de padrões [57]. Assim, a PCA foi utilizada para explicar as principais diferenças entre os *clusters* formados com o uso do HCA.

A quantidade de informação extraída após a aplicação de uma PCA depende do número de componentes principais (PCs) consideradas. Os dois métodos comumente utilizados para determinar o número de componentes são o teste de *elbow*, ou *scree*, e a seleção das componentes com autovalor maior que 1 [57]. Observando a Figura 8, é possível notar que o *elbow* não está nitidamente definido nas primeiras componentes do *scree plot*. Além disso, existem 68 PCs com autovalor maior que 1.

Tendo em vista esses aspectos, optou-se por selecionar um conjunto de componentes com autovalor maior que 1 e que explicassem no mínimo 90% da variância dos dados. Desse modo, 23 PCs (90,27% da variância) foram selecionadas e o HCA foi realizado sobre essas componentes.

Figura 8 – Gráficos *Scree plot,* e gráfico Autorvalor x PC.



A interpretação de uma PCA é realizada por meio da inspeção dos gráficos dos escores e dos pesos. O primeiro corresponde às coordenadas dos objetos no espaço das PCs, enquanto o segundo fornece a relação entre as PCs e as variáveis, permitindo explicar a distribuição dos objetos no gráfico dos escores [58].

O gráfico dos escores (PC1 X PC2), com 6 *clusters* formados, está disposto na Figura 9. Enquanto a Figura 10 apresenta o gráfico de pesos das 2393 variáveis, o qual pode ser utilizado para descrever a disposição dos corantes em cada *cluster*.

Verifica-se que a PC1 tem contribuições majoritárias dos descritores estruturais e alguns quânticos, como o potencial de ionização (P.I.) e dureza (H). Por outro lado, a PC2 traz as informações referentes aos espectros de absorção, eletronegatividade (E.N.) e eletropositividade (E.P.). Também é visível que as cargas parciais pouco contribuem para a separação dos indivíduos.

Figura 9 – Gráfico dos escores projetados em PC1 (32,1%) X PC2 (12,2%) com os *clusters* formados.

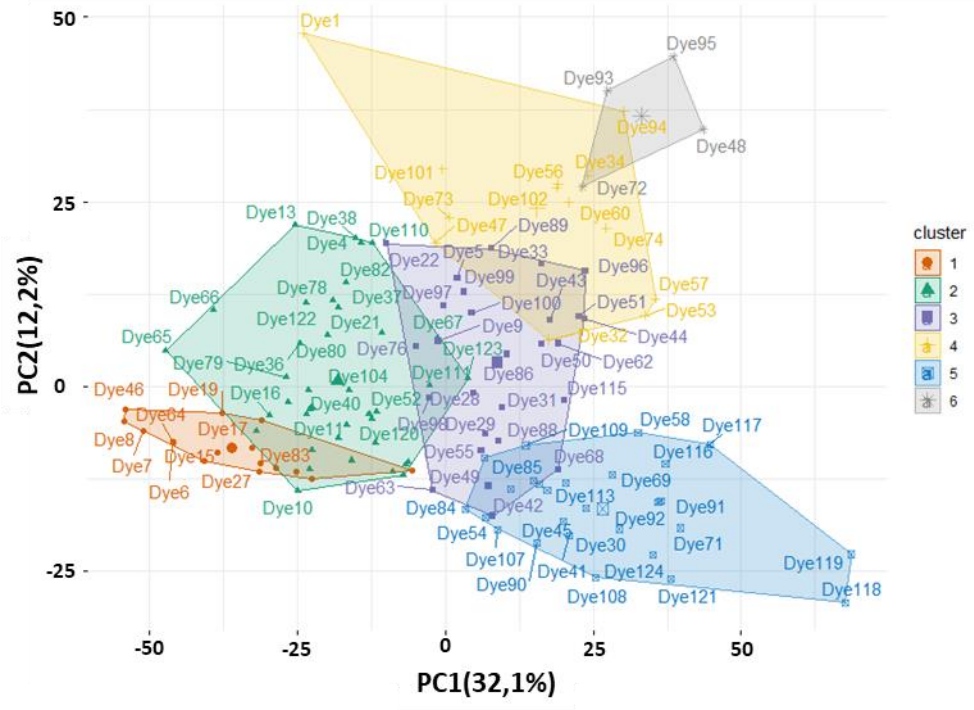
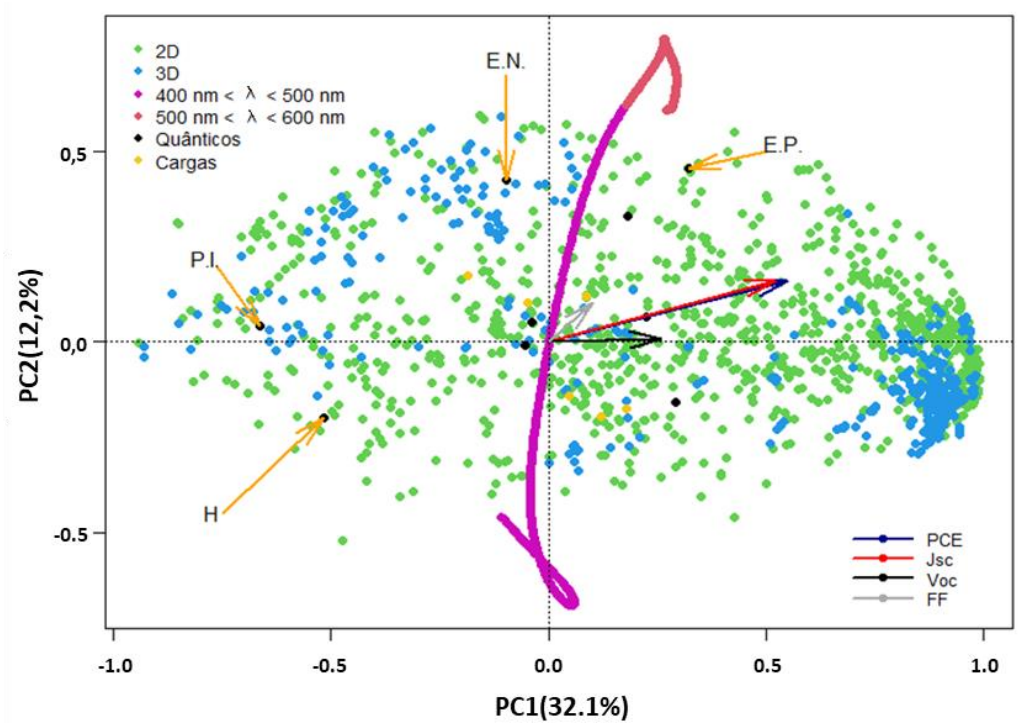


Figura 10 – Gráfico dos pesos nas duas primeiras componentes: PC1 (32,1%) x PC2 (12,2%).



Uma inspeção nas contribuições dos descritores para a PC1 revela que os corantes mais à direita dessa componente tendem a ser mais volumosos, com uma maior maciez, um menor potencial de ionização e tendem a ser os corantes

que compõem as células de maiores PCE. Para as espécies mais à esquerda, o inverso é observado.

Em contrapartida, os corantes que apresentam os maiores valores na PC2 tendem a absorver em comprimentos de onda cada vez mais próximos de 600 nm. Dessa maneira, é possível atribuir algumas características gerais e comparações entre os *clusters* formados.

Por exemplo, os *clusters* 1 e 2 apresentam corantes menos volumosos, mais duros e com maiores potenciais de ionização que aqueles do *cluster* 5. Além disso, é possível salientar que as espécies dos *clusters* 4 e 6 tendem a possuir absorções máximas em comprimentos de onda próximos de 600 nm, diferente dos corantes dos *clusters* 1 e 5 que tem suas máximas absorções próximas de 400 nm. Já o *cluster* 3 contém estruturas que apresentam características intermediárias dos outros grupos.

Dessa maneira, após as amostragens aleatórias em cada um dos *clusters*, o grupo de treinamento (95 estruturas) e o de teste (31 estruturas) foram formados. Vale salientar que o *Dye56* e *Dye59*, respectivamente o PCE máximo e mínimo, foram forçadamente incluídos no grupo de treino. A Tabela 5 apresenta a distribuição dos corantes de cada *cluster* nos grupos de treino e teste.

Tabela 5 – Distribuição dos corantes nos grupos de treino e teste.

Cluster	**Corantes**	
	Treino	**Teste**
1	8, 14, 17, 19, 20, 23, 24, 25, 26, 27, 46, 83	6, 7, 15, 64
2	2, 3, 4, 10, 11, 12, 13, 16, 35, 37, 38, 40, 52, **59**, 66, 67, 77, 79, 81, 82, 87, 103, 104, 105, 110, 114, 120, 122, 123, 126	18, 21, 36, 39, 65, 75, 78, 80, 111
3	28, 29, 31, 42, 43, 49, 50, 51, 55, 62, 63, 68, 76, 86, 89, 96, 97, 98, 100, 115	5, 9, 22, 33, 44, 88, 99
4	1, 32, 34, 53, **56**, 60, 73, 74, 94, 102	47, 57, 101
5	30, 41, 45, 54, 58, 61, 69, 70, 71, 84, 85, 90, 91, 109, 112, 116, 118, 119, 124, 125	92, 106, 107, 108, 113, 117, 121
6	48, 93, 95	72

59 – PCE mínimo; **56 –** PCE máximo

4.3 Parametrização do GA

Como descrito na seção 3.4, buscou-se um conjunto de parâmetros para o algoritmo genético que retornasse os modelos mais preditivos para o conjunto de treino. Para isso, o tamanho da população, número de gerações, probabilidade de mutação e tipo de cruzamento foram explorados. No apêndice 8.3 estão apresentados os gráficos referentes às modificações.

Observando como o método de cruzamento afeta a predição dos modelos, percebe-se que independentemente dos outros parâmetros, o método *single* e *random* resultam valores de R^2 aproximadamente idênticos para os melhores modelos. Porém, o método *single* fornece um conjunto de modelos com coeficientes de determinação médio maior que o *random*. Desse modo, optou-se por definir o método de cruzamento como *single*.

Realizar a seleção de variáveis por GA utilizando um grande número de indivíduos e de gerações tende a aumentar o custo computacional envolvido na otimização. Em vista disso, como não há um aumento substancial nos coeficientes de determinação a partir de 500 gerações, seja do melhor modelo ou do conjunto obtido, e que 250 indivíduos geram, em média, modelos tão bons quanto os gerados por 350 indivíduos, optou-se por selecionar 500 gerações e 250 indivíduos.

Por fim, observando como a modificação da probabilidade de mutação para os parâmetros já selecionados afetam os modelos obtidos, observa-se que uma taxa de 0,05 proporciona os maiores valores de R^2 para os melhores modelos.

Assim, o algoritmo genético operou com a seguinte configuração: Tamanho da População = 250; Número de Gerações = 500; Mínimo e Máximo de Variáveis de 5 e 12, respectivamente; Probabilidade de Mutação = 0,05 e *Crossover* = *single.*

4.4 Modelos para o PCE

A partir da execução do *script* utilizando o algoritmo genético e as métricas de corte (R^2_{Treino} ≥ 0,6, R^2_{Pred}≥ 0,5 e, para cada descritor, valor-p ≤ 0,05 e VIF ≤ 5), foi possível obter 198 modelos, porém apenas os 15 com maiores R^2_{Pred} foram

selecionados para a etapa de refinamento. Dessa maneira, após empregar o *best-subsets* em conjunto com os novos critérios de escolha (R^2_{Treino}≥ 0,7, R^2_{Pred} ≥ 0,65, valor-p ≤ 0,05 e VIF ≤ 5, para cada descritor e no máximo 14 variáveis), apenas 4 modelos foram obtidos.

Nas Tabelas 6 – 9 estão apresentados os coeficientes dos modelos e seus respectivos desvios padrões, valores p e VIF. Verifica-se que todos os 4 modelos obtidos para a predição do PCE satisfazem as seguintes condições [81]:

I. ***n*** > *4* ***k****,* sendo ***n*** o número de compostos no conjunto de treinamento e ***k*** o número de descritores no modelo;
II. Todos os descritores são estatisticamente significantes, pois p < 0,05;
III. Todos os descritores são fraca/moderadamente correlacionados, já que VIF ≤ 5.

Tabela 6 – Coeficientes do modelo 1 (M-1) - ***k*** = 10, F = 23,14.

Coeficientes	**Valor**	**σ**	**Valor-p**	**VIF**
Constante	41,46292	7,054022	$8{,}14\cdot10^{-8}$	-
ATSC4e	1,462918	0,324847	$2{,}14\cdot10^{-5}$	2,628949
ATSC4s	-0,12125	0,017594	$9{,}50\cdot10^{-10}$	2,067891
GATS7c	-5,8755	0,866361	$1{,}55\cdot10^{-9}$	1,567342
VE2_Dzi	-98,97	26,66648	$3{,}7\cdot10^{-4}$	1,45624
C1SP3	0,161461	0,065133	0,015177	2,809662
minHBint8	-0,98589	0,194341	$2{,}31\cdot10^{-6}$	2,003224
MDEO-11	-5,71084	2,585898	0,029936	1,408635
PNSA-3	0,050951	0,016882	0,003368	2,217062
Carea	0,006561	0,001062	$2{,}25\cdot10^{-8}$	3,937671
SpMax2_Bhp	-6,50814	1,707646	$2{,}63\cdot10^{-4}$	2,726352

Tabela 7 – Coeficientes do modelo 2 (M-2) - ***k*** = 12, F = 30,92.

Coeficientes	**Valor**	**σ**	**Valor-p**	**VIF**
Constante	51,16831	7,122512	$2{,}81\cdot10^{-10}$	-
ATSC4c	2,803665	0,910097	0,002811	1,269299
ATSC4s	-0,11229	0,014665	$3{,}33\cdot10^{-11}$	2,063361
GATS7c	-4,21276	0,670592	$1{,}52\cdot10^{-8}$	1,348756
GATS5p	-4,53076	1,224726	$3{,}9\cdot10^{-4}$	1,493726
VE1_Dzs	-1,26471	0,305386	$8{,}34\cdot10^{-5}$	1,251768
SpMin2_Bhm	-7,36176	2,68316	0,00746	1,649342
minHBint8	-1,37174	0,179734	$3{,}72\cdot10^{-11}$	2,460994
FNSA-2	5,09554	0,828361	$2{,}67\cdot10^{-8}$	2,909385
Cvol	0,007484	0,000642	$4{,}22\cdot10^{-19}$	3,736921
SpMax2_Bhp	-4,61821	1,50938	0,002994	3,059384
$I_{540.681363}$	1,209957	0,554061	0,031834	1,388099
elea	-0,88621	0,308555	0,005188	1,572026

Tabela 8 – Coeficientes do modelo 3 (M-3) - ***k*** = 13, F = 26,85.

Coeficientes	Valor	σ	Valor p	VIF
Constante	39,82667	6,769271	$8{,}68\cdot10^{-8}$	-
ALogP	-0,3491	0,059892	$1{,}09\cdot10^{-7}$	3,887371
ATSC4e	1,412073	0,285838	$4{,}13\cdot10^{-6}$	2,775173
ATSC4s	-0,1221	0,016521	$1{,}18\cdot10^{-10}$	2,485841
GATS7c	-5,33477	0,904342	$8{,}12\cdot10^{-8}$	2,328411
VE2_Dzi	-84,1936	23,53078	$5{,}88\cdot10^{-4}$	1,545974
SpMax2_Bhp	-5,87094	1,610034	$4{,}69\cdot10^{-4}$	3,304335
C1SP3	0,229488	0,054312	$6{,}21\cdot10^{-5}$	2,663656
minHBint8	-1,21327	0,176489	$1{,}18\cdot10^{-9}$	2,252507
MDEO-11	-11,4577	2,511115	$1{,}77\cdot10^{-5}$	1,811074
Carea	0,004125	0,000941	$3{,}45\cdot10^{-5}$	4,208833
ATSC7c	-2,13864	0,83853	0,012642	1,615318
BCUTc-1l	-5,39563	2,270046	0,019818	1,073549
RotBFrac	-8,28164	2,394473	$8{,}68\cdot10^{-4}$	4,5432

Tabela 9 – Coeficientes do modelo 4 (M-4) - ***k*** = 14, F = 31,34.

Coeficientes	Valor	σ	Valor p	VIF
Constante	37,65727	4,617922	$4{,}03\cdot10^{-12}$	-
ATS6m	0,000234	1,87E-05	$1{,}81\cdot10^{-20}$	3,54903
ATSC6i	0,020144	0,004756	$6{,}06\cdot10^{-5}$	1,720361
MATS4s	-22,955	3,737759	$2{,}99\cdot10^{-8}$	1,763323
GATS7c	-4,41269	0,752539	$9{,}72\cdot10^{-8}$	1,944914
VE2_Dzv	-61,1769	20,56018	0,003867	1,563843
BCUTc-1l	-10,9212	2,183905	$3{,}31\cdot10^{-6}$	1,198583
SpMin1_Bhi	-12,6455	2,168057	$1{,}11\cdot10^{-7}$	2,548464
VE3_Dt	0,046871	0,019064	0,016106	2,00148
ndO	-1,05362	0,252099	$7{,}41\cdot10^{-5}$	1,612155
FNSA-2	5,691081	0,774961	$1{,}54\cdot10^{-10}$	2,915722
AATSC5s	4,710096	2,159336	0,0321	1,941682
SaasC	-0,09164	0,024867	$4{,}15\cdot10^{-4}$	2,082705
minHBint8	-1,01721	0,179446	$2{,}20\cdot10^{-7}$	2,808934
$I_{434.468938}$	0,83896	0,420679	0,049526	1,179178

Como já mencionado, observar como os resíduos estão distribuídos em função das predições permite que algumas conclusões sejam feitas, tais como incerteza na obtenção da constante do modelo, erro de análise dos dados e a homoscedasticidade [57,62]. Dessa maneira, os gráficos de resíduos *vs* valores preditos e o normal QQ dos resíduos padronizados serão apresentados para os quatro modelos selecionados.

Na Figura 11 estão os gráficos dos resíduos *vs* valores preditos para todos os modelos. Nota-se que os resíduos estão aleatoriamente distribuídos entre va-

lores de |ê| ≤ 2,5, indicando que os resíduos são independentes e estão distribuídos em torno do ê = 0. Os gráficos normal QQ dos resíduos padronizados (Figura 12) mostram que os resíduos se distribuem normalmente, porém observa-se a existência de pequenos desvios da normalidade nas caudas direita e esquerda.

Figura 11 – *Gráficos dos resíduos vs valores preditos.*

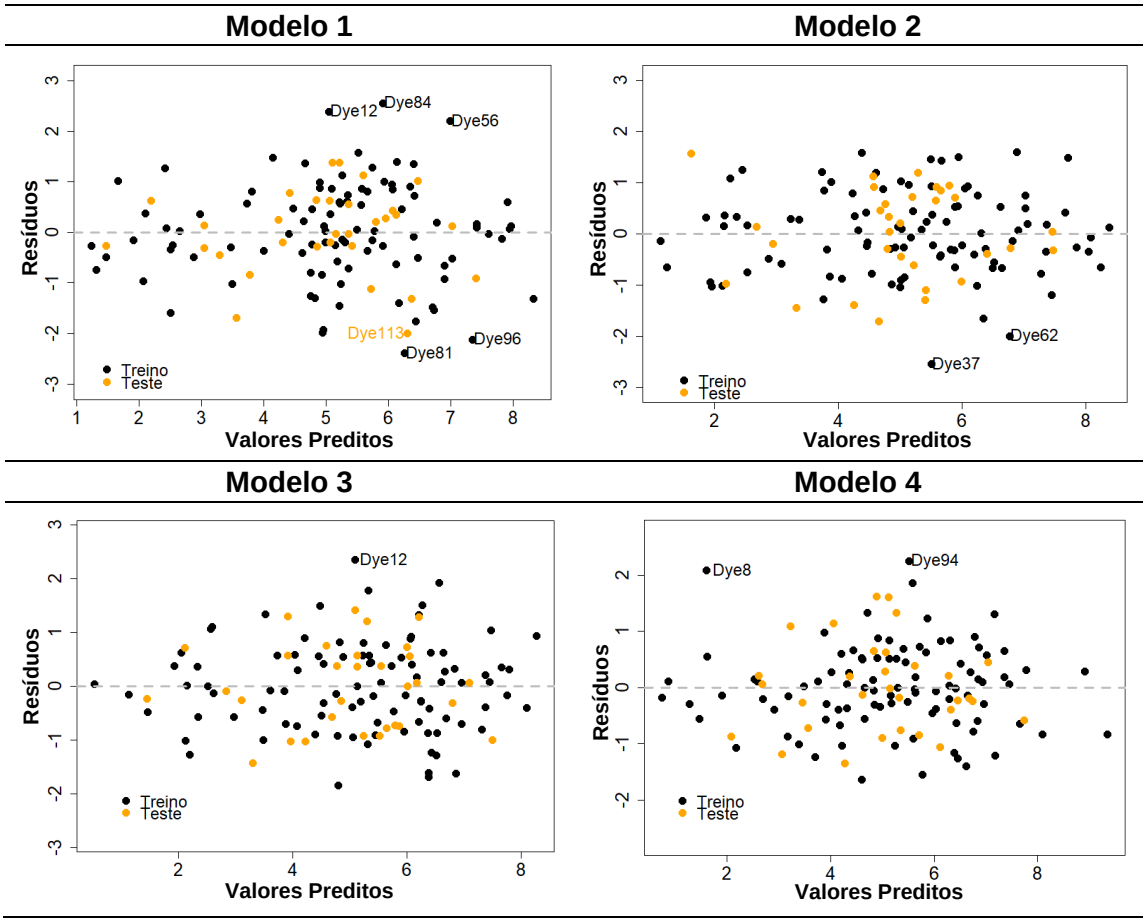
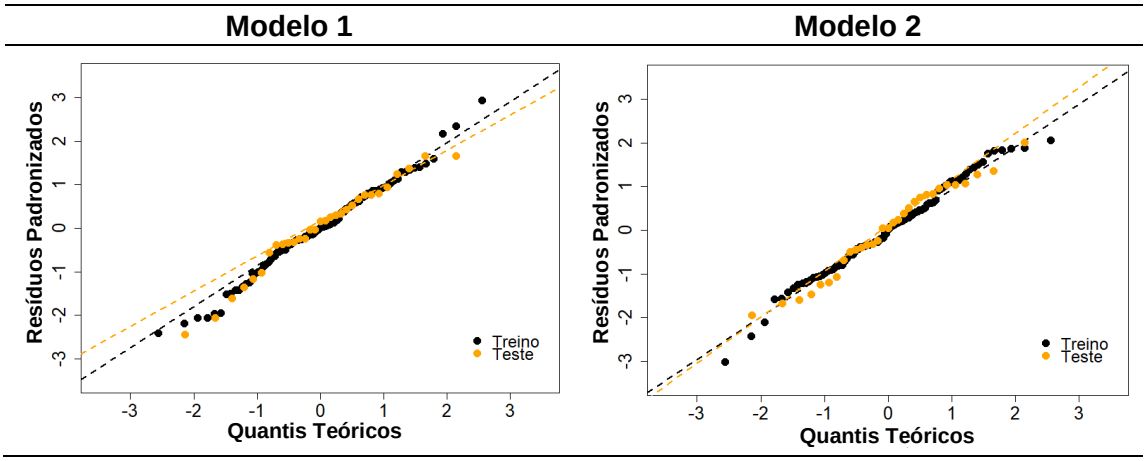
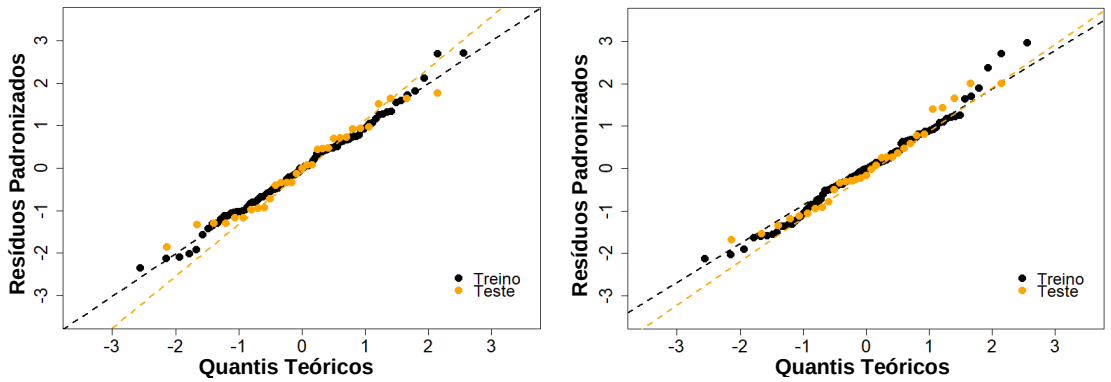


Figura 12 – *Normal QQ dos resíduos padronizados.*



Modelo 3	Modelo 4



Para cada um dos modelos apresentados, as métricas de ajuste (R^2_{Treino} e R^2_{Adj}), de previsibilidade interna (Q^2_{LOO}, obtida por meio da validação cruzada *leave-one-out*) e medidas de erro (RMSE e MAE) estão apresentadas na Tabela 10.

Tabela 10 – Métricas de variância explicada pelo modelo, previsibilidade interna e medidas de erro.

Modelo	R^2_{Treino}	R^2_{Adj}	Q^2_{LOO}	RMSE	RMSEcv	MAE	MAEcv
M-1	0,734	0,702	0,630	0,986	1,162	0,771	0,891
M-2	0,819	0,792	0,753	0,812	0,949	0,652	0,762
M-3	0,812	0,781	0,711	0,829	1,026	0,668	0,804
M-4	0,846	0,819	0,781	0,750	0,894	0,577	0,692

Nota-se que os modelos obtidos são capazes de explicar uma quantidade considerável da variabilidade dos dados, já que todos apresentam R^2_{Treino}> 0,7. Além disso, ao analisar os valores de R^2_{Adj}, o qual é mais punitivo ao aumento da complexidade do modelo, têm-se um indicativo de que o número de descritores utilizados não conduz ao *overfitting*, tendo em vista a similaridade entre R^2_{Treino} e R^2_{Adj} [66]. Somado a isso, os valores de Q^2_{LOO} > 0,6 reforçam a ausência do *overfitting* e indicam uma boa previsibilidade interna dos modelos [66,67], a qual é corroborada pela semelhança entre as medidas de erro do ajuste e da validação cruzada.

Quando os modelos são comparados baseados nos coeficientes de determinação e medidas de erro, percebe-se que o M-1 apresenta o conjunto de métricas inferior, enquanto o M-4 e M-2 apresentam os melhores ajustes e previsibilidade interna, pois esses apresentam os maiores R^2 e Q^2_{LOO} e os menores

valores de RMSE e MAE. Já o M-3 possui um ajuste e capacidade de previsão muito similar aos M-4 e M-2.

Baseado na regressão através da origem e na validação cruzada *leave-one-out*, Roy e colaboradores [69] propuseram duas métricas ($\overline{r^2_{m\,(LOO)}}$ e $\Delta r^2_{m\,(LOO)}$) para validação interna de modelos QSPR. A Tabela 11 apresenta os termos necessários para a obtenção dessas métricas e os critérios que um dado modelo deve satisfazer para ser considerado internamente validado.

Tabela 11 – Métricas r^2_m para validação interna.

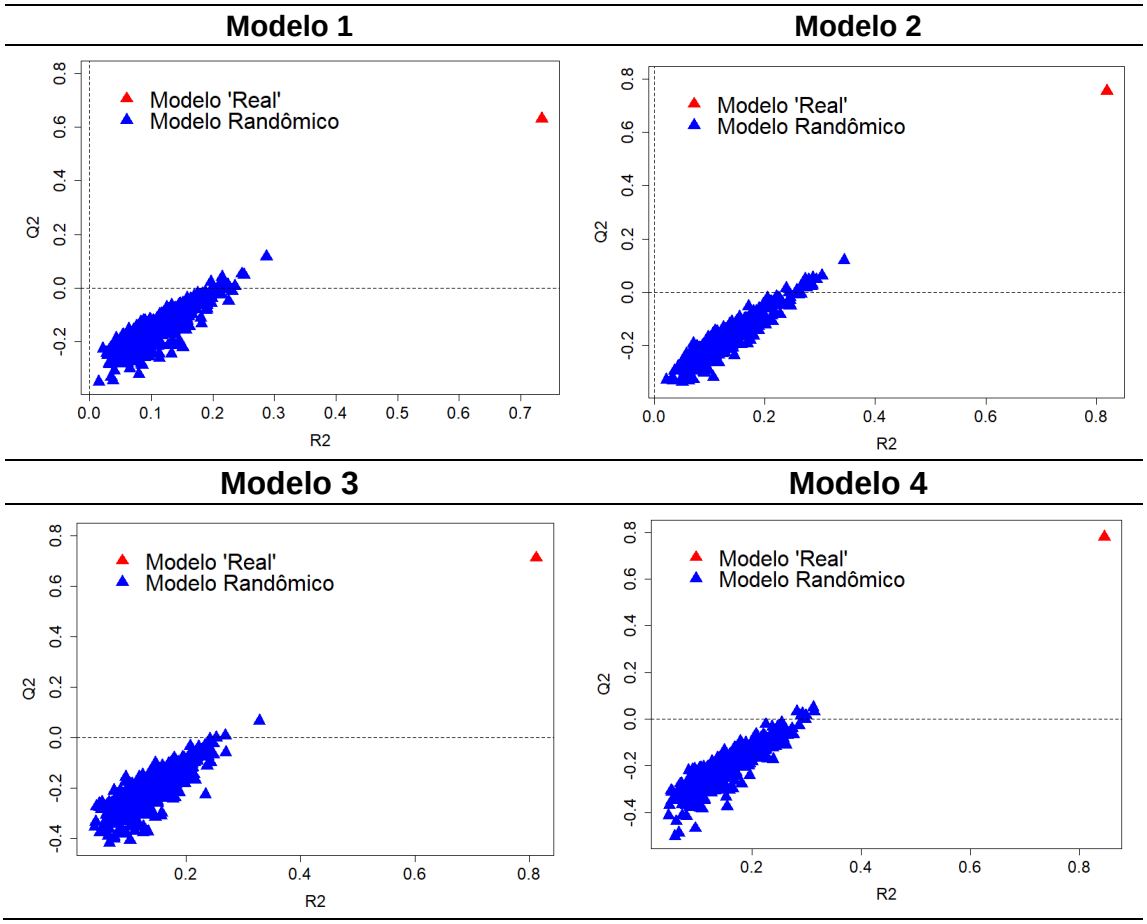
		Modelo			
Métrica	**Critério**	**M-1**	**M-2**	**M-3**	**M-4**
$\boldsymbol{r^2_{(LOO)}}$	-	0,633	0,735	0,714	0,782
$\boldsymbol{R^2_{o(LOO)}}$	-	0,631	0,734	0,712	0,781
$\boldsymbol{R'^2_{o(LOO)}}$	-	0,512	0,673	0,648	0,746
$\boldsymbol{r^2_{m(LOO)}}$	-	0,601	0,713	0,682	0,759
$\boldsymbol{r'^2_{m(LOO)}}$	-	0,411	0,553	0,531	0,632
$\overline{\boldsymbol{r^2_{m\,(LOO)}}}$	> 0,5	0,506	0,633	0,607	0,696
$\Delta\boldsymbol{r^2_{m\,(LOO)}}$	< 0,2	0,189	0,160	0,151	0,126

Nota-se que todos os modelos satisfazem as condições propostas por Roy e colaboradores [69], assim é possível dizer que os modelos obtidos apresentam uma satisfatória capacidade preditiva interna. Para Roy *el al.* [69], um modelo é mais preditivo internamente quanto mais próximo de 1 for o $\overline{r^2_{m\,(LOO)}}$ e mais próximo de 0 for o $\Delta r^2_{m\,(LOO)}$, reforçando a ordem decrescente de melhores modelos: M-4 > M-2 > M-3 > M-1.

Como já mencionado, o *y-radomization* foi performado para avaliar a significância de cada um dos quatro modelos apresentados. Para isso, a cada iteração, um modelo obtido e suas métricas R^2 e Q^2_{LOO} foram registradas e comparadas com as do modelo 'real'. A Figura 13 apresenta os gráficos $Q^2_{LOO}\ X\ R^2$ para as 500 iterações de cada um dos modelos.

Observa-se que em todas as randomizações realizadas, os modelos randômicos apresentaram baixos valores de R^2 , os quais são, em grande parte, acompanhados por valores negativos de Q^2_{LOO}. Como os coeficientes de determinação dos modelos 'reais' destoam fortemente daqueles randômicos, pode-se afirmar que os descritores selecionados nos quatro modelos têm uma conexão com o PCE.

Figura 13 – Gráficos *das randomizações do vetor resposta.*



Até o momento, os modelos apresentados foram internamente validados por meio de métricas de ajuste e robustez. Porém, é necessário observar a capacidade de prever o PCE para estruturas que não estão presentes no treinamento. Ou seja, faz-se necessário o processo de validação externa, onde um conjunto de objetos que não foi utilizado no treinamento é adotado para avaliar a real capacidade preditiva do modelo.

Sabe-se que um baixo Q^2 indica a baixa previsibilidade de um modelo, porém não é possível dizer que um alto valor de Q^2 esteja associado com uma forte capacidade de prever a propriedade de compostos não utilizados no treinamento [69]. Por esse motivo, algumas métricas são utilizadas para avaliar tal capacidade, uma comumente utilizada é obtida conforme a equação 27, segundo a qual um modelo pode ser descrito como preditivo se $R^2_{Pred} \geq 0{,}5$ [71]. A Tabela 12 apresenta o R^2_{Pred} e algumas medidas de erro para a validação externa.

Tabela 12 – Coeficiente de determinação, RMSE e MAE para o conjunto teste.

Modelo	R^2_{Pred}	$RMSE_{pred}$	MAE_{pred}
M-1	0,709	0,858	0,694
M-2	0,681	0,899	0,744
M-3	0,709	0,859	0,709
M-4	0,743	0,807	0,659

Dos quatro modelos, M-4 e M-2 apresentam a melhor e a pior capacidade de previsão, respectivamente, do conjunto apresentado. Já M-1 e M-3 possuem uma previsibilidade muito similar, tendo em vista os valores de R^2_{Pred}, $RMSE_{pred}$ e MAE_{pred}. Nota-se ainda, que as medidas de erro para o grupo de treino (Tabela 10) e as de teste (Tabela 12) são, em geral, muito parecidas e consideravelmente próximas de zero.

Assim como os procedimentos de validação interna, a validação externa normalmente adota outras métricas além do R^2_{Pred}, o que reforça as conclusões acerca da capacidade de predição de um dado modelo. Por exemplo, a regressão através da origem aplicada sobre o conjunto de teste resulta em um conjunto de parâmetros que podem ser utilizados para validar externamente um modelo.

Neste estudo foram adotadas as métricas propostas por Roy *et al.* [69] e Golbraikh e Tropsha [67], as quais estão apresentadas na Tabela 13 juntamente com seus os valores mínimos aceitáveis. Observando a tabela, nota-se que todos os modelos propostos estão de acordo com os critérios estabelecidos, e que a tendência do poder de predição já apresentada (M-4 > M-3 ≈ M-1 > M-2) é reforçada.

Tabela 13 – Métricas de extras para validação externa.

		Modelo			
Métrica	**Critério**	**M-1**	**M-2**	**M-3**	**M-4**
r^2	$> 0{,}6$	0,732	0,700	0,757	**0,752**
R_o^2	-	0,706	0,697	0,719	**0,747**
r_m^2	$> 0{,}5$	0,614	0,658	0,611	**0,698**
k	$0{,}85 \leq k \leq 1{,}15$	1,003	0,959	0,964	**0,975**
$\frac{r^2 - R_o^2}{r^2}$	$< 0{,}1$	0,037	0,005	0,053	**0,007**

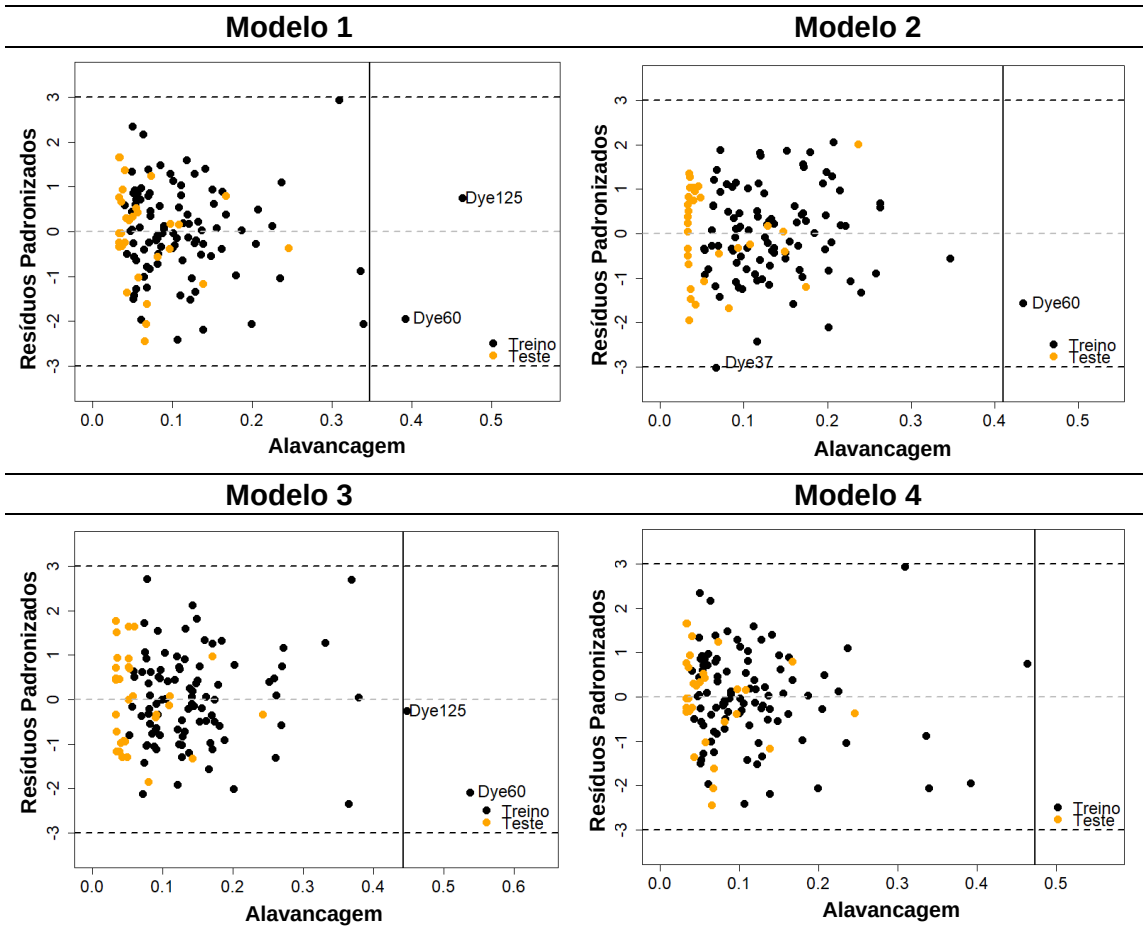
Como a similaridade da capacidade preditiva entre M-1 e M-3 é reforçada por essas métricas, é possível escolher apenas um deles com base na complexidade dos mesmos. Além da semelhança da capacidade de predição, nota-se que eles apresentam um conjunto de descritores muito similares, nove dos dez descritores usado no M-1 estão presentes no M-3. Desse modo, é possível afirmar que os descritores que diferenciam o M-3 e M-1, só afetam o ajuste do conjunto de treinamento e pouco influenciam no poder de predição. Nesse sentido, o M-1 é preferível, já que ele apresenta o menor número de descritores.

Além de avaliar a previsibilidade, a OECD recomenda que, ao desenvolver um modelo QSAR/QSPR, deve-se definir o Domínio de Aplicabilidade (DA) e verificar se os compostos em estudo estão compreendidos no mesmo [52]. Um método frequentemente utilizado para definir o DA é verificar se os valores de alavancagem de cada um dos objetos dos conjuntos de treino e teste são menores de $h^* = 3p/n$, sendo n o número de objetos no conjunto de treinamento e p o número de descritores mais 1 [61]. Comumente, um gráfico de resíduos padronizados em função da alavancagem (gráfico de Willams) é plotado para observar tanto o DA, quanto os possíveis *outliers* [52,61].

A Figura 14 apresenta os gráficos de Willams para todos os quatro modelos. Nota-se que todos os objetos do grupo de treinamento e teste apresentam resíduos padronizados localizados entre ±3 em M-1, M-3 e M-4, indicando a ausência de *outliers*. Todavia, no M-2 o *Dye32* possui um resíduo padronizado menor que o limite inferior (-3,013), porém como esse corante tem uma baixa influência no modelo e sua distância de Cook é menor que 1 (Apêndice 8.4), ele foi mantido no conjunto de dados.

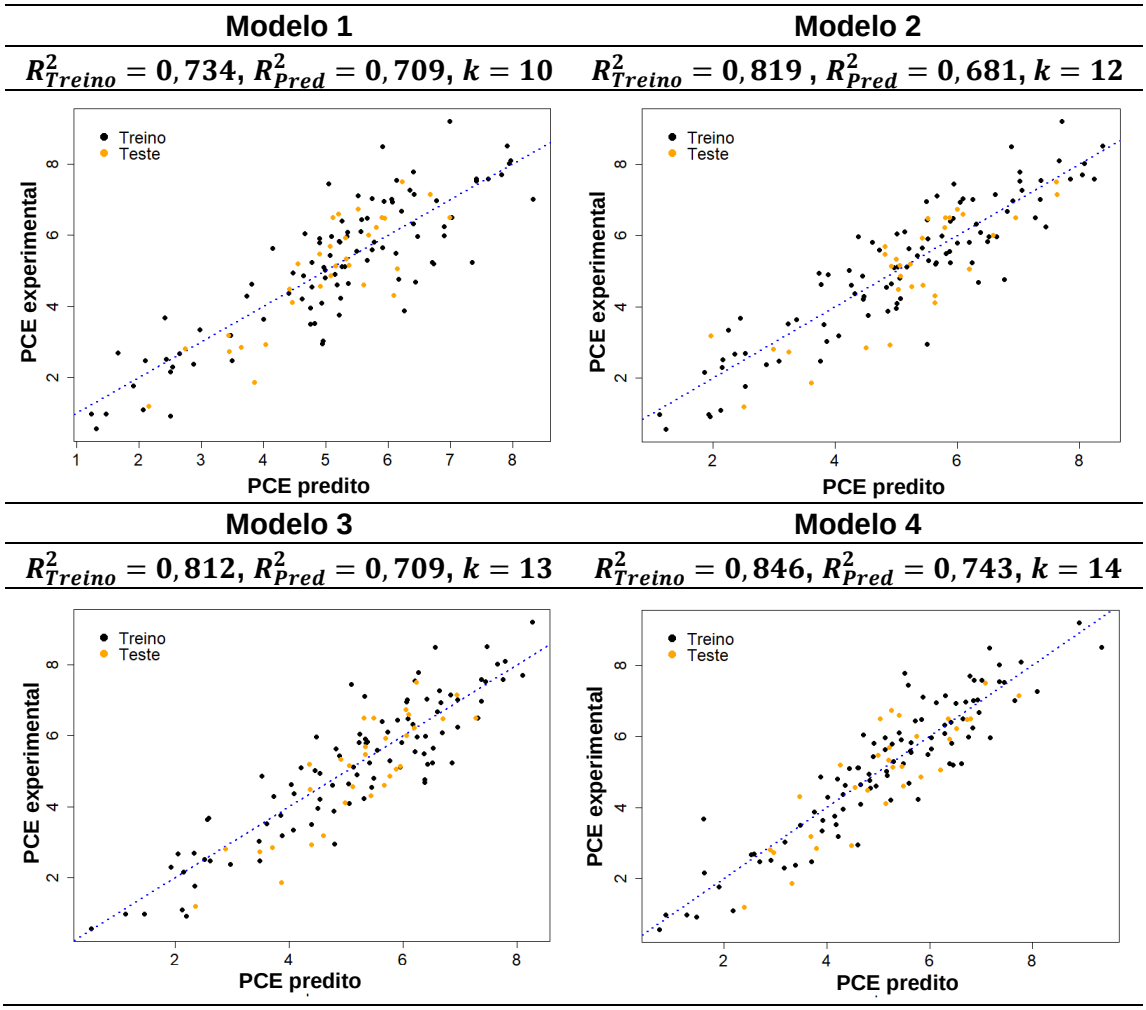
Além disso, é perceptível que apenas no M-4 os objetos não apresentaram valores de alavancagem superior ao valor limite (h^* = 0,4736). Nos modelos M-1 (h^* = 0,3474), M-2 (h^* = 0,4105) e M-3 (h^* = 0.4421), o *Dye60* está fora do DA, indicando que previsão da propriedade para esse corante utilizando tais modelos não é confiável, pois o modelo é extrapolado para obtenção de tal medida. Somado a isso, percebe-se que os descritores que distinguem o M-1 do M-3 não fazem com que os corantes *Dye125* e *Dye60* sejam inseridos no DA, reforçando a preferência pelo M-1.

Figura 14 – Gráfico de Willams. A linha sólida representa o valor limite de h* e o pontilhado os resíduos padronizados ±3.



Por fim, a Figura 15 apresenta os gráficos de PCE *vs.* PCE predito para os modelos apresentados. A partir de uma inspeção visual, nota-se que os modelos apresentados possuem um ajuste e uma capacidade de predição suficientemente razoáveis, tendo em vista a proximidade dos pontos com a reta $\boldsymbol{PCE = PCE_{Predito}}$.

Figura 15 – Gráficos PCE vs. PCE predito.



4.4.1 Interpretação do modelo M-4 para o PCE

Dos modelos apresentados anteriormente, o M-4 foi o que apresentou o maior poder de ajuste e predição para a eficiência de conversão de uma DSSC. Desse modo, essa seção é dedicada a possível interpretação dos descritores presentes nesse modelo.

O M-4 apresenta descritores que podem ser divididos em 5 tipos diferentes: de Autocorrelação; Autovalor/autovetor; Eletrotopológico; CPSA e Espectroscópico. Todos esses descritores estão apresentados no Quadro 1, juntamente suas influências – Positiva (+) ou negativa (-) – para o PCE de uma dada DSSC.

Quadro 1 – Descritores do M-4 e suas descrições.

Descritor	Influência	Descrição	Tipo
ATS6m	+	Autocorrelação Broto-Moreau– lag 6 ponderada por massa	Autocorrelação
ATSC6i	+	Autocorrelação Broto-Moreau centrada – lag 6 ponderada pelo primeiro potencial de ionização	
AATSC5s	+	Autocorrelação Broto-Moreau centrada média – lag 5 ponderada por I-State	
MATS4s	-	Autocorrelação Moran – lag 4 poderada por I-*State*	
GATS7c	-	Autocorrelação de Geary – lag 7 ponderada por cargas	
VE2_Dzv	-	Média dos coeficientes do último autovetor da matriz de Barysz / ponderada por volumes de van der Waals	Autovalor/autovetor
SpMin1_Bhi	-	Valor absoluto do menor autovalor da matriz de Burden modficada – n 1 / ponderada pelo primeiro potencial de ionização	
BCUTc-1l	-	Menor autovalor ponderado por cargas parciais/BCUTs	
VE3_Dt	+	Logaritmo da soma dos coeficientes do último autovetor da matriz detour	
minHBint8	-	Mínimo *E-State* da força para potenciais ligações de hidrogênio em caminhos de comprimento 8	Eletrotopológico
ndO	-	Número de átomos com E-State: =O	
SaasC	-	Soma dos átomos de *E-State:* :C:-	
FNSA-2	+	PNSA-2 / área da superfície molecular	CPSA
$I_{434.468938}$	+	Absorção no visível	Espectroscópico

o $I_{434.468938}$ (+) é facilmente interpretável, pois indica que corantes que apresentam absorções mais intensas próximas de 434 nm tendem a proporcionar maiores PCE. Tal descritor reforça a ideia de que os corantes devem absorver mais intensamente na região do visível para possibilitar maiores eficiências de conversão [4,5].

Outra característica é apresentada pelo descritor FNSA-2 (do inglês *"Fractional negative surface area"*), o qual faz parte do conjunto de descritores CPSA (do inglês *"Charged partial surface area"*) proposto por Stanton e Jurs [82]. Por definição, o FNSA-2 é um valor negativo dado por:

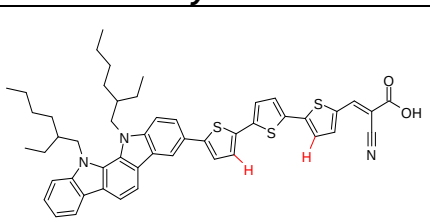
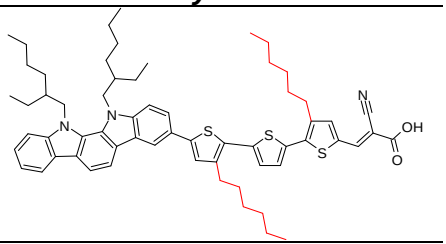
$$FNSA-2=\frac{PNSA-2}{SASA}=\frac{Q^{-}\sum_{a^{-}}SA_{a}^{-}}{SASA} \qquad \text{(Eq. 28)}$$

Sendo Q^{-} a soma de todas as cargas parciais negativas na molécula, SA_a^- corresponde à área superficial negativa no átomo a que é acessível ao solvente e $SASA$ descreve a área superficial acessível ao solvente. Nota-se que o termo $\sum_{a^-} SA_a^-/SASA$ é um valor positivo que corresponde a fração molecular que apresenta carga parcial negativa e é acessível ao solvente. Ou seja, o FNSA-2 descreve a fração negativa da molécula que o solvente pode interagir, porém ponderada pela carga negativa total.

Observando a influência do FNSA-2 (+), e sabendo que ele sempre assume um valor negativo, espera-se que o aumento do valor absoluto de FNSA-2 proporcione a redução da eficiência da célula. A partir da Eq. 28, nota-se que a modificação da área acessível ao solvente pode proporcionar a redução do valor absoluto de FNSA-2. Por exemplo, a inserção de cadeias alquílicas pode resultar no acréscimo da $SASA$, modificando pouco a distribuição de cargas negativas.

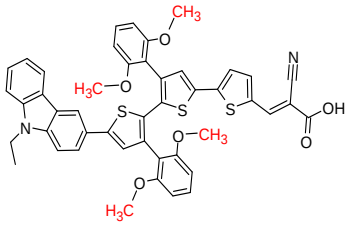
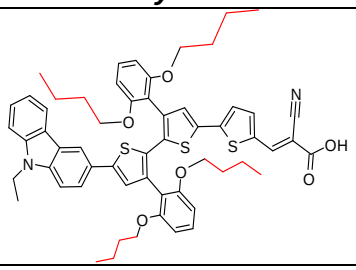
As Tabelas 14 e 15 apresentam algumas comparações onde a inserção/aumento de cadeias alquílicas é acompanhada pelo aumento do PCE. A presença de cadeias laterais (comumente as alquílicas) em sensibilizadores é reportada na literatura, pois elas normalmente auxiliam na redução das interações π-π, as quais tendem a afetar negativamente o PCE [45].

Tabela 14 – Comparação entre *Dye29* e *Dye30* para a inserção de cadeias alquílicas na ponte π.

	Dye29	***Dye30***
		
$\sum_{a^-} SA_a^-/SASA$	0,37	0,29
$FNSA-2$	-1,09	-1,04
SaasC	10,63	14,42

Se as modificações estruturais passarem pela inserção de cadeias alquílicas, o modelo M-4 apresenta um indicativo das posições que potencialmente terão um efeito negativo sobre o valor de PCE. Lembrando que deve haver cautela na inserção de cadeias alquílicas, pois o aumento excessivo do volume molecular resultará na redução da quantidade de corante adsorvido na superfície do semicondutor e, consequentemente, no valor de PCE [83,84].

Tabela 15 – Comparação entre *Dye73* e *Dye74* para o aumento das cadeias alquílicas nos substituintes da ponte π.

	Dye73	***Dye74***
		
$\sum_{a^-} SA_a^-/SASA$	0,34	0,27
$FNSA-2$	-1,22	-1,08

A informação sobre as posições vem da influência do descritor SaasC (-), o qual faz parte do conjunto de descritores eletrotopológicos propostos por Kier e Hall [85,86]. Esses descritores são capazes de computar características elétricas de um átomo considerando o efeito da vizinhança sobre ele [85,86]. Assim, é possível distinguir um carbono de anel aromático daqueles de cadeias alquílicas ou identificar se um átomo é mais interno ou periférico em uma molécula, com base no valor de *E-State, S,* proposto por esses autores. O *E-State* de um dado átomo é calculado segundo:

$$S_i = I_i + \Delta I_i = \underbrace{\frac{\left(2/\mathrm{N}\right)^2 \delta^v + 1}{\delta}}_{I_i} + \underbrace{\sum \frac{I_i - I_j}{r_{ij}^2}}_{\Delta I_i} \qquad \text{(Eq. 29)}$$

O valor I_i descreve o valor intrínseco, ou *I-State,* do átomo *i* e ΔI_i corresponde a modificação do valor intrínseco devido à presença dos outros átomos na molécula. *N* corresponde ao número quântico principal do nível de valência do átomo, r_{ij} a distância entre o átomo i e j, δ^v e δ correspondem aos valores de conectividade molecular, os quais são obtidos da seguinte maneira:

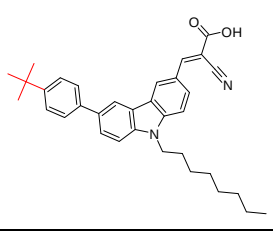
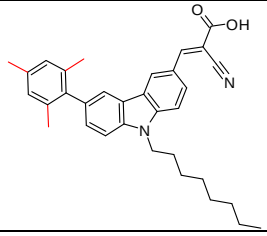
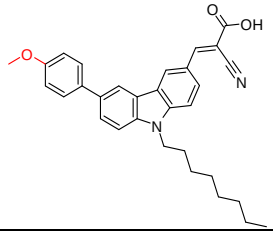
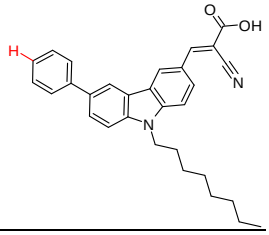
$$\delta = \sigma - H \quad \text{(Eq. 30)}$$

$$\delta^{v} = \sigma + \pi + n - H \quad \text{(Eq. 31)}$$

Onde σ, π, *n* o número de elétrons em orbitais σ, π e em pares isolados, respectivamente; e H corresponde ao número de hidrogênios ligados ao dado átomo.

O SaasC descreve a soma dos valores de *E-State* de todos os carbonos aromáticos substituídos presentes na molécula, ou seja, substituir anéis aromáticos tende a diminuir o PCE de uma DSSC. Esse efeito pode ser observado com a inserção de cadeias alquílicas na ponte π do *Dye29* (Tabela 14) e com a substituição em *Dye23, Dye24, Dye25* (Tabela 16).

Tabela 16 – Modificação do SaasC por substituições nos anéis aromáticos.

	Dye23	***Dye24***	***Dye25***	***Dye26***
				
SaasC	4,35	6,95	3,73	3,03

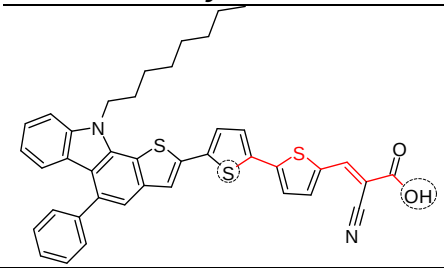
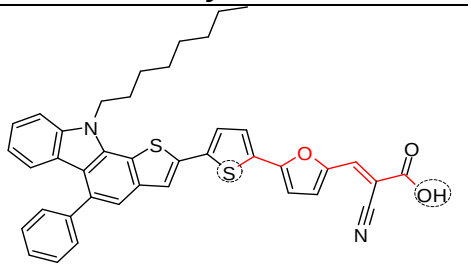
Outra característica apontada pelo modelo M-4 é dada pela influência de outro descritor eletrotopológico proposto por Kier e Hal [85,86], o minHBint8 (-). Esse descritor é obtido pelo menor valor de *E-state* dentre os átomos que potencialmente formam ligações de hidrogênio intramoleculares, sendo que o átomo receptor da ligação e o hidrogênio estão à 8 ligações de distância. Nas estruturas utilizadas nesse estudo, existe no máximo 1 átomo que satisfaz a condição definida pelo descritor.

Como esse descritor apresenta uma influência negativa, espera-se que a redução do minHBint8 proporcione maiores eficiências de conversão da DSSC. A Tabela 17 apresenta uma comparação entre dois corantes com diferentes valores de minHBint8, salientando que o nHBint8 corresponde ao número de potenciais ligações de hidrogênio intramolecular à 8 ligações.

Sabe-se que o aumento da eletronegatividade de um dado átomo em uma da cadeia tende a proporcionar a redução dos *E-State* de todos os outros átomos

na molécula [86]. Assim, é possível reduzir o valor do minHBint8 inserindo átomos mais eletronegativos em determinadas posições na molécula, lembrando que a perturbação no valor intrínseco depende do quadrado da distância entre os átomos (Eq. 29).

Tabela 17 – Comparação dos valores de minHBint8 para *Dye21* e *Dye22*.

	Dye21	***Dye22***
		
nHBint8	1	1
minHBint8	-0,88	-1,04

O último dos descritores eletrotopológicos é o ndO (-), o qual contabiliza a presença de oxigênios com ligações duplas. Como todas as estruturas apresentam um grupo carboxila, possivelmente a influência negativa do ndO sugere que o aumento no número oxigênio duplamente ligado tende a reduzir a eficiência da célula.

Os descritores de autocorrelação (ATS6m, ATSC6i, AATSC5s, MATS4s e GATS7c) indicam como uma determinada propriedade (eletronegatividade, carga, massa, volume etc) está distribuída através da molécula composta por *A* átomos [48]. Genericamente, o descritor de correlação pode ser obtido segundo a expressão:

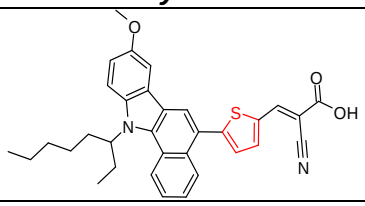
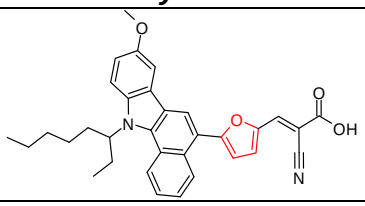
$$D5_{(\tau;\sigma,\lambda,k)} = \alpha \sum_{i=1}^{A} \sum_{j=1}^{A} \left(\tau_i \cdot \tau_j\right)_{ij}^{\lambda} \cdot \delta\left(d_{ij};k\right) \qquad \text{(Eq. 32)}$$

Sendo τ um valor genérico invariante referente à propriedade atômica; α corresponde ao termo para escalar os dados, λ é o expoente do termo $(\tau_i.\tau_j)_{ij}$; δ (d_{ij}; k) é a função delta de Kronecker, a qual assume valor 1 quando d_{ij} = k e 0 para $d_{ij} \neq k$; O termo *k* (o lag) corresponde a distância topológica entre o átomo *i* e *j* [48]. Para cada tipo de autocorrelação, a Eq. 32 assume formas características.

O ATS6m (+) aparenta ter uma correlação positiva com a massa molecular, indicando que o aumento da massa tende a aumentar o valor de ATS6m, contribuindo para a elevação do PCE. Como, no geral, a modificação da massa molecular tende a ser acompanhada pela alteração do volume molecular, pode-se relacionar o significado desse descritor com o aumento do volume molecular.

Já o ATSC6i (+) parece indicar que a presença de átomos com a primeira energia de ionização mais altas tendem a favorece o aumento da eficiência de conversão de uma DSSC. Indicando que a presença de átomos menos susceptíveis a perda de elétrons tende a proporcionar maiores PCEs. A Tabela 18 apresenta dois corantes que se diferenciam apenas pela ponte π – anel tiofênico (*Dye39*) e furânico (*Dye40*) – e apresentam diferentes valores de ATS6m e ATSC6i.

Tabela 18 – Comparação entre os valores do ATSC6i para o Dye39 e Dye40.

	Dye39	***Dye40***
		
ATS6m	14219,40	13367,14
ATSC6i	-35,52	-26,96

O MATS4s (-) é o descritor de autocorrelação de Moran para os *I-State* a 4 ligações, esse descritor comumente apresenta valores no intervalo [-1, +1], onde valores próximos de zero indicam que não há um padrão específico na distribuição da propriedade atômica na distância considerada [48]. O MATS4s apresenta um valor médio para as 126 estruturas de 0,15±0,08, o que indica que não há um forte padrão na distribuição dos *I-State* dos átomos a 4 ligações.

O AATSC5s (+) também é um descritor de autocorrelação ponderado pelo *I-State*, porém com a influência positiva. Diferente do MATS4s, o AATSC5s parece sofrer mais fortemente com a modificação dos valores intrínsecos dos átomos presentes na molécula. As Tabelas 20 e 21 apresentam duas comparações entre os valores de AATSC5s e MAT4s para dois pares de corantes consideravelmente similares e a Tabela 19 contém os *I-States* de alguns tipos de átomos.

Tabela 19 – *I-State* de alguns átomos.

Tipo de átomo	*I-State*	Tipo de átomo	*I-State*
>C=	1,67	=CH-	2,00
-O-	3,50	-S-	1,83
=O	7,00	=S	3,67

Tabela 20 – Comparação entre os valores de AATSC5s e MAT4s para o Dye44 e Dye45.

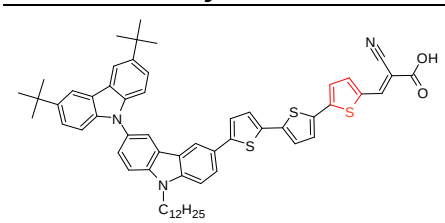
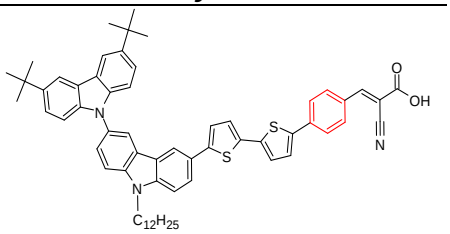
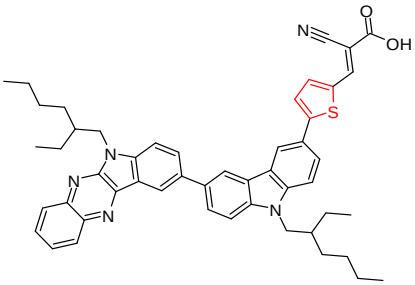
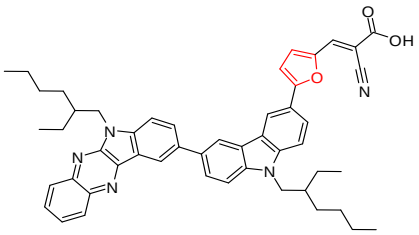
	Dye44	***Dye45***
		
AATSC5s	$3{,}45.10^{-3}$	$5{,}44.10^{-2}$
MAT4s	0,164	0,167

Tabela 21 – Comparação entre os valores de AATSC5s e MAT4s para o Dye54 e Dye55.

	Dye54	***Dye55***
		
AATSC5s	-0,0151	0,1069
MAT4s	0,1343	0,1349

Também parece haver um padrão relacionado à distribuição de cargas parciais, a qual é indicado pelo descritor GATS7c (-). Esse é o descritor de Geary para as cargas parciais atômicas à 7 ligações que indica uma autocorrelação positiva quando exibe um valor entre 0 e 1. Como o GATS7c, para os 126 corantes, apresenta um valor médio de 0,82±0,08, tem-se um indicativo de que há uma tendência de aumento nas cargas parciais para átomos relativamente distantes. Como esse descritor apresenta uma influência negativa no modelo M-4, espera-se que o PCE aumente com a diminuição da diferença entre as cargas parciais dos átomos relativamente distantes na molécula.

Até o momento, já foram apresentados os possíveis significados de apenas 10 descritores, porém ainda restam quatro: os baseados em autovalor/autovetor. Esses descritores são obtidos a partir de operações específicas realizadas sobre os autovalores e/ou coeficientes dos autovetores de uma dada matriz *M.*

O modelo M-4 apresenta quatro descritores desse tipo, os quais são obtidos a partir da matriz de Barysz (VE2_Dzv), de Burden (SpMin1_Bhi e BCUTc-1l) e de detour (VE3_Dt). Na Tabela 22 estão apresentadas as definições dessas matrizes[48,87–89].

Tabela 22 – Definição das matrizes de Barysz, Burden e detour.

Matriz	Definição	
Barysz	$\left[D\right]_{ij} = \begin{cases} d_{ij}\left(w,\ \pi^*\right) = \sum_{b=1}^{d_{ij}} \left(\frac{1}{\pi_b^*} \cdot \frac{w_C^2}{w_{b(1)} \cdot w_{b(2)}} \right), \text{ se } i \neq j \\ 1\text{-}\frac{w_C}{w_i} \quad , \text{ se } i = j \end{cases}$	(Eq. 33)
Burden	$\left[B\right]_{ij} = \begin{cases} \pi_{ij}.10^{-1}, \text{ se } (i,\ j) \in \mathrm{E}(G) \\ w_i \quad , \text{ se } i = j \\ 0,001 \ , \text{ se } (i,\ j) \notin \mathrm{E}(G) \end{cases}$	(Eq. 34)
detour	$\left[\Delta\right]_{ij} = \begin{cases} \Delta_{ij} = \left\lvert {}^{\max} p_{ij} \right\rvert, \text{ se } i \neq j \\ 0 \quad , \text{ se } i = j \end{cases}$	(Eq. 35)

Nas expressões acima, w_C e w_i correspondem, respectivamente, a uma dada propriedade (volume de van der Waals, potencial de ionização, polarizabilidade etc) do átomo de carbono e do i-ésimo átomo; π refere-se a ordem de ligação entre o par de átomos (1 para simples, 2 para dupla, 3 para tripla e 1,5 para aromática); $^{max}p_{ij}$ corresponde à maior distância topológica entre os átomos *i* e *j*.

O VE2_Dzv (-) corresponde à média dos coeficientes do último autovetor da matriz de Barysz ponderada por volumes de van de Waals (vdW). A matriz de Barysz foi inicialmente proposta utilizando os números atômicos como peso, porém foi posteriormente generalizada para qualquer propriedade atômica, w [48]. Como tal DM é obtido a partir da matriz ponderada pelo volume de vdW, espera-se que a modificação no volume de um grupo de átomos resultará na alteração

do VE2_Dzv.

Observando a alteração do VE2_Dzv com determinadas mudanças estruturais, percebe-se que, no geral, o aumento no número de anéis aromáticos presentes na ponte π conduzem ao abaixamento do valor de VE_Dzv (Tabela 23). Além disso, aparentemente, a inserção de cadeias alquílicas nas pontes ou doadores também produzem um abaixamento desse DM (Tabela 24), a qual corrobora com a redução do valor absoluto de FNSA-2 e aumento do ATS6m. Devido a influência do VE2_Dzv (-), espera-se que a redução do seu valor proporcione o aumento do PCE da DSSC.

Tabela 23 – Modificação do VE2_Dzv com o aumento da ponte π.

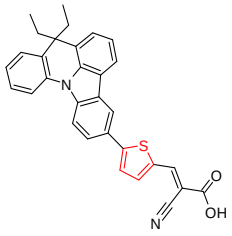
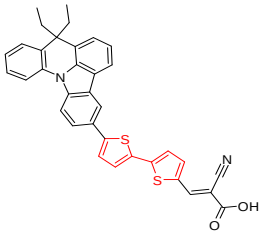
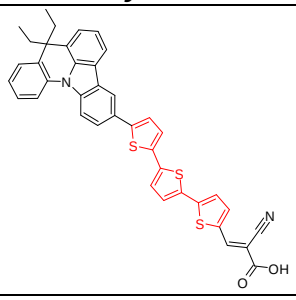
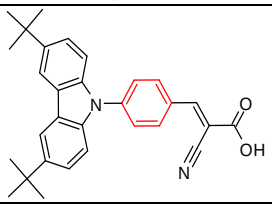
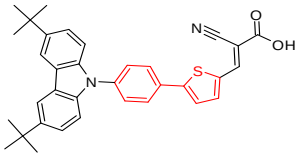
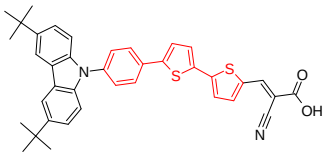
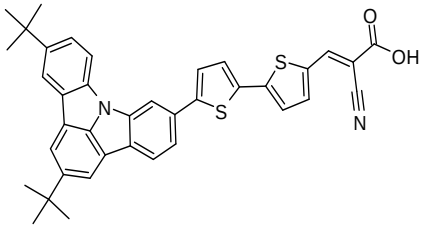
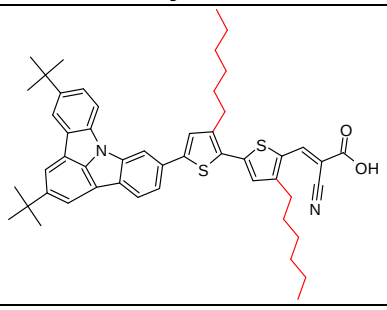
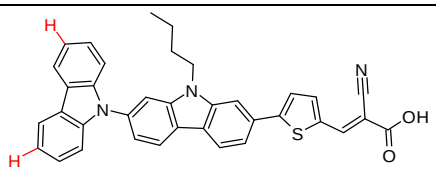
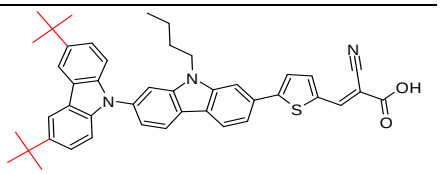
	Dye35	***Dye36***	***Dye37***
			
VE2_Dzv	$9{,}8.10^{-3}$	$8{,}4.10^{-3}$	$7{,}0.10^{-3}$
	Dye14	***Dye10***	***Dye11***
			
VE2_Dzv	$9{,}9.10^{-3}$	$7{,}6.10^{-3}$	$5{,}7.10^{-3}$

Tabela 24 – Modificação do VE2_Dzv com a inserção de cadeias alquílicas.

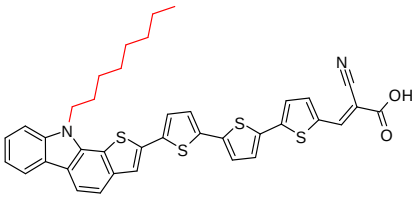
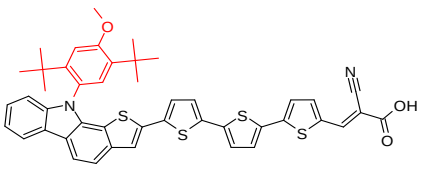
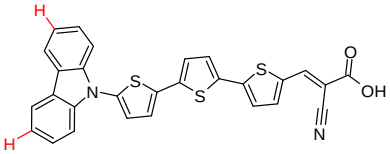
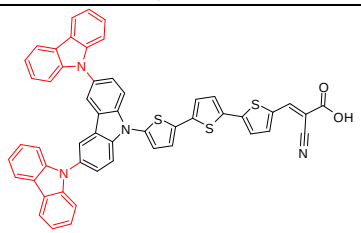
	Dye87	***Dye88***
		
VE2_Dzv	$7{,}7.10^{-3}$	$5{,}1.10^{-3}$
	Dye2	***Dye3***

		
VE2_Dzv	$5{,}4.10^{-3}$	$4{,}3.10^{-3}$

O SpMin1_Bhi (-) corresponde ao valor absoluto do menor autovalor da matriz de Burden ponderada pelo primeiro potencial de ionização. O conjunto dos menores autovalores dessa matriz têm sido apontados como descritores capazes de discriminar – e proporcionar o ordenamento das – estruturas moleculares [48].

Desse conjunto de menores autovalores, o menor contém a contribuição de todos os átomos, o que reflete a característica geral da molécula [48]. Então, como a propriedade atômica observada é o potencial de ionização, espera-se que a alteração do valor do SpMin1_Bhi tenha uma considerável relação com a modificação no potencial de ionização da molécula. A Tabela 25 apresenta algumas comparações dos valores de SpMin1_Bhi com a modificação dos substituintes no grupo doador, onde nota-se que o abaixamento do valor do SpMin1_Bhi é acompanhado com o aumento do potencial de ionização da molécula.

Tabela 25 – Modificação do SpMin1_Bhi com a alteração do grupo substituinte no doador.

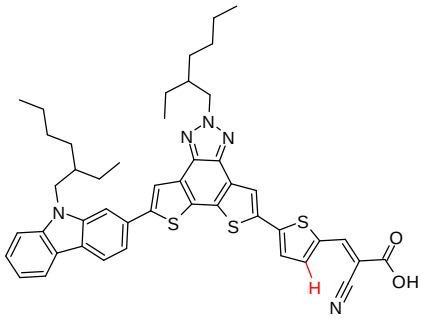
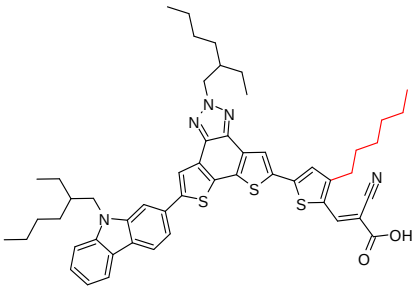
	Dye110	***Dye111***
		
SpMin1_Bhi	1,97	2,02
P.I.	8,00	7,94
	Dye66	***Dye67***
		
SpMin1_Bhi	1,94	2,03
P.I.	8,18	7,87

No caso do BCUTc_1l, o qual corresponde ao primeiro menor autovalor da matriz de burden ponderada por cargas parciais, provavelmente, ele reforça o indicativo envolvendo cargas atômicas já apresentadas pelo descritor GATS7c.

O VE3_Dt (+) corresponde ao logaritmo da soma dos coeficientes do último autovetor da matriz detour. Como essa matriz é construída baseada apenas na maior distância topológica que separa dois vértices em uma estrutura, espera-se que moléculas essencialmente idênticas apresentem o mesmo valor de VE3_Dt. Estruturas similares com valores iguais de VE3_Dt são os *Dye39* e *Dye40* (Tabela 18), onde esse descritor tem o valor -9.34, e os *Dye54* e *Dye55* (Tabela 21), cujo descritor vale -5.80.

Além dessa característica, esse descritor aparenta diferenciar estruturas que apresentam um aumento na ramificação. A Tabela 26 apresenta algumas dessas observações. Assim, tendo em vista e influência do VE3_Dt (+) pode-se esperar que modificações que aumentem a ramificação da estrutura do corante proporcione o aumento do PCE de uma DSSC.

Tabela 26 – Modificação do VE3_Dt com o aumento da ramificação.

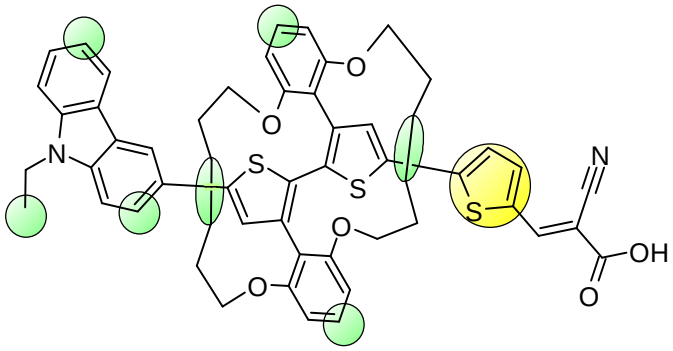
	Dye89	***Dye90***
		
VE3_Dt	-8,67	-6,28

4.5 Modificações estruturais e predição da performance fotovoltaica

Do conjunto de 126 corantes selecionados, o *Dye56* é o que apresenta a maior eficiência de conversão. Por esse motivo, ele foi escolhido para a realização de modificações estruturais visando o aumento do PCE.

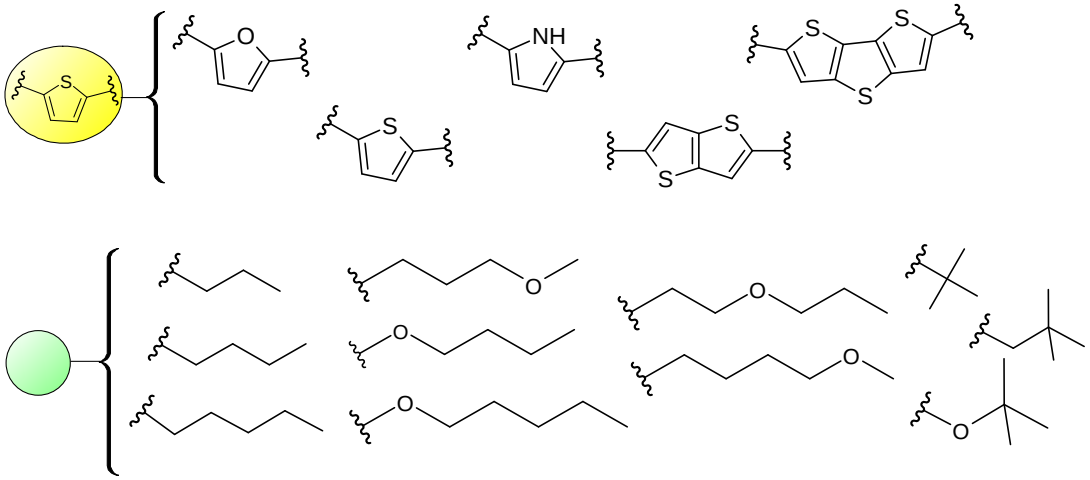
As regiões em que as modificações estruturais foram realizadas estão apresentadas na Figura 16.

Figura 16 – Pontos de modificações no Dye56. Os círculos indicam as regiões que soferam modificação.



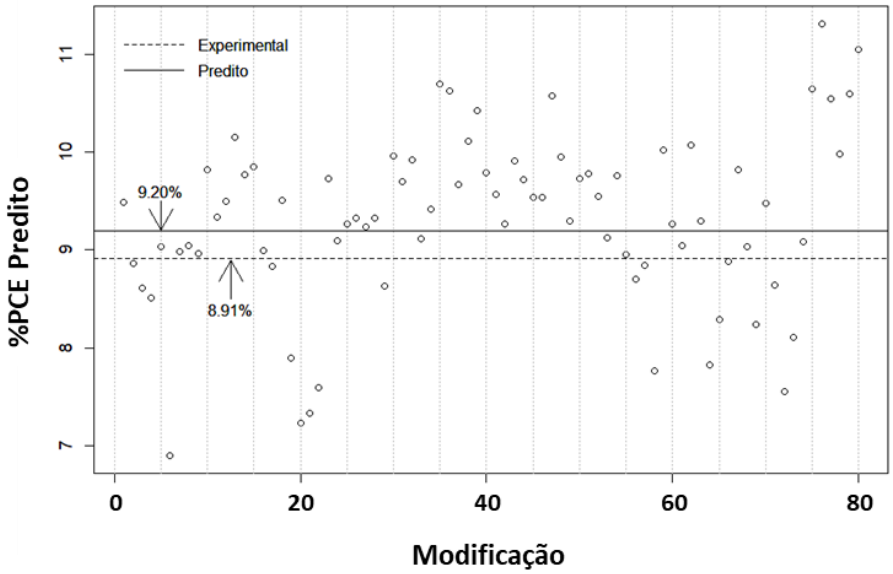
Na realização das modificações, procurou-se seguir as características moleculares indicadas pelo modelo M-4. De uma maneira geral, buscou-se: (i) Aumentar a ramificação molecular; (ii) modificação da ponte π; (iii) Inserir átomos com maiores potenciais de ionização. A Figura 17 apresenta os grupos utilizados nas modificações.

Figura 17 – Grupos utilizados para nas modificações.



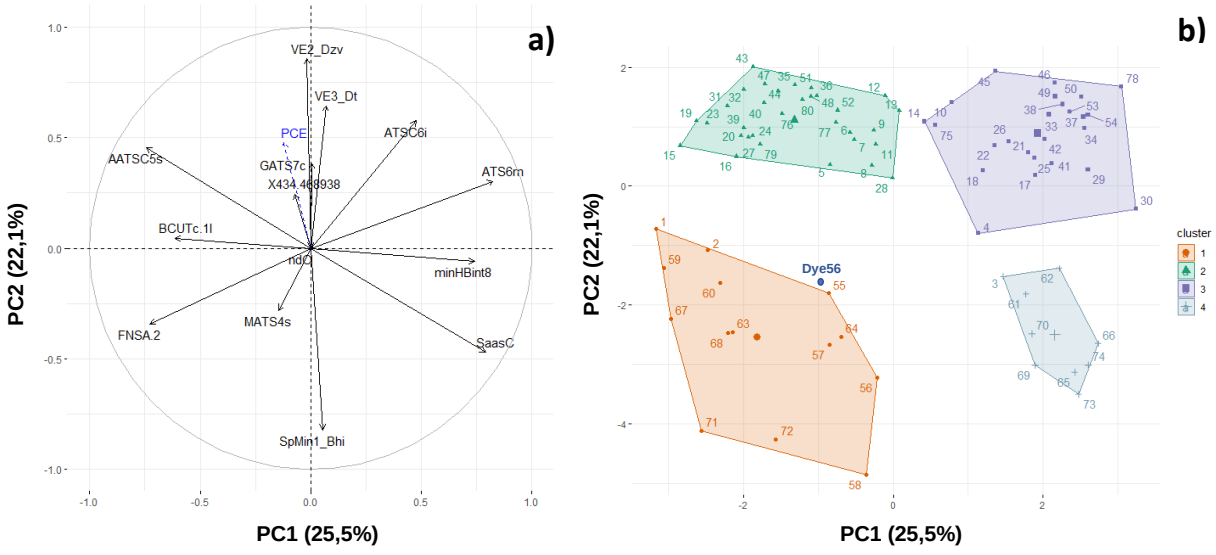
Dessa maneira, foi possível obter um total de 80 modificações estruturais (Apêndice 8.2). Todas essas estruturas foram otimizadas e seus descritores moleculares obtidos seguindo a metodologia apresentada na seção 3.2. Por fim, o modelo M-4 foi empregado para predizer a eficiência de conversão de cada uma das modificações realizadas. O Figura 18 apresenta a eficiência de conversão em função das modificações.

Figura 18 – Dispersão %PCE Predito vs Modificações. A linha sólida destaca o PCE experimental do *Dye56*; A linha tracejada destaca o PCE predito pelo modelo M-4.



Na tentativa de identificar padrões presentes nas modificações realizadas, um agrupamento hierárquico combinado com a análise de componentes principais foi feito (seguindo o procedimento apresentado na seção 3.3). Para isso, apenas os 14 descritores presentes no modelo M-4 e as componentes com autovalor maior que 1 foram selecionadas (apenas 5 componentes, explicando cerca de 75% da variância). A Figura 19 apresenta os gráficos de pesos e escores com os *clusters* formados.

Figura 19 – Gráficos projetados nas duas primeiras componentes para as 80 modificações: a) pesos; b) escores com os clusters formados. O PCE e o *Dye56* foram tratados como variável e indivíduo suplementares, respectivamente.

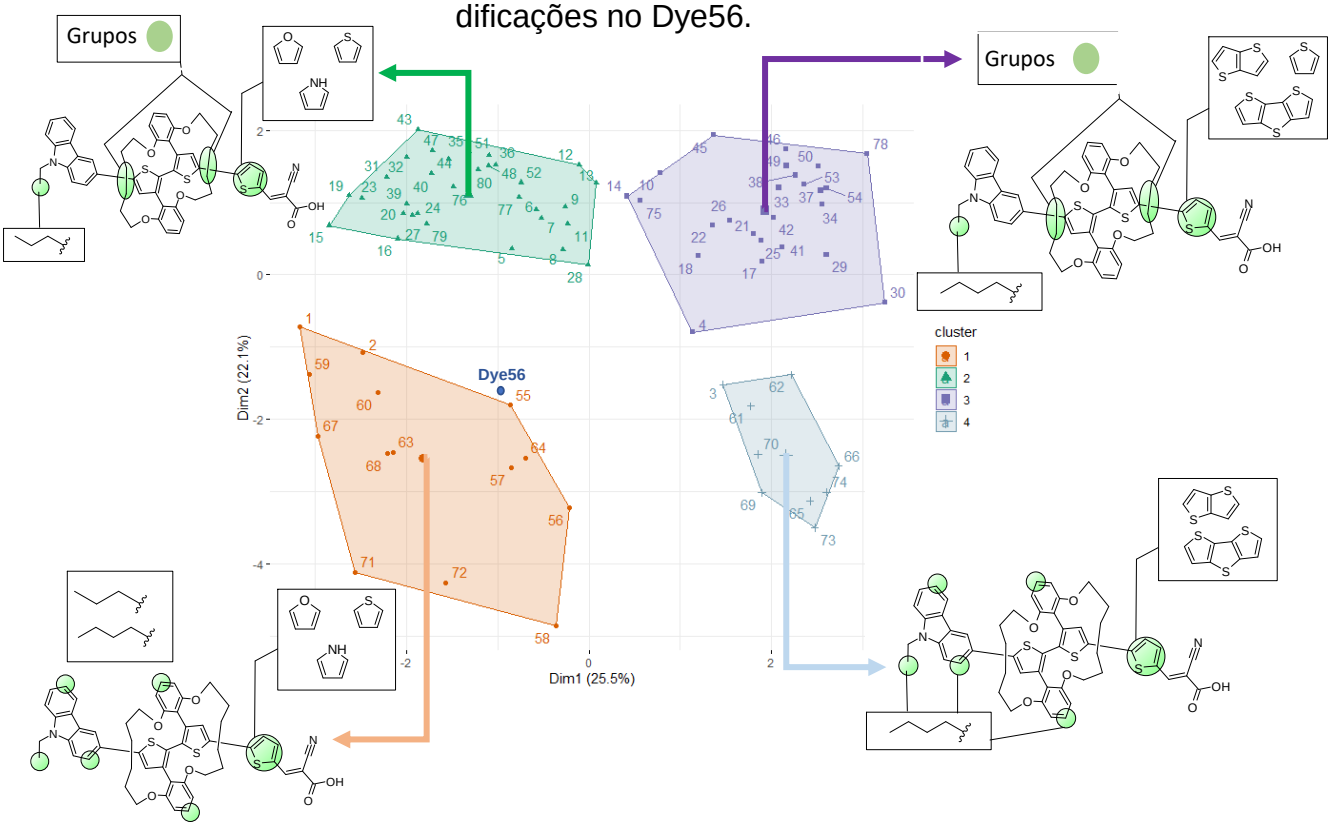


No gráfico dos pesos (Figura 19-a), nota-se que a primeira componente apresenta maiores contribuições dos descritores SaasC, minHBint8, ATS6m, AA-TSC5s, BCUTc-1l e FNSA-2. Enquanto os descritores SpMin_Bhi, VE2_Dzv e VE3_Dt contribuem mais fortemente para a composição da segunda componente. Apesar das duas primeiras componentes explicarem cerca de 48% da variância dos dados, ainda é possível observar uma satisfatória distinção entre os quatro *clusters* formados (Figura 19-b).

Por exemplo, os *clusters* 1 e 2 têm, no geral, os maiores valores de AA-TSC5s, BCUTc-1l e FNSA-2. Por outro lado, os *clusters* 3 e 4 possuem maiores valores de SaasC, minHBint8, ATS6m. Além disso, nota-se que os grupos com maiores valores na PC2 apresentam maiores valores de VE2_Dzv e VE3_Dt, enquanto os grupos com menores valores nessa componente tendem a apresentar maiores valores de Spmin1_Bhi.

Ainda na Figura 19-a, é possível destacar que os grupos com maiores valores na PC2 tentem a apresentar um aumento no PCE, destacando os *clusters* 2 e 3. Já na Figura 19-b, nota-se que resultaram no cluster 1 ainda são bastante similares à estrutura de origem, *Dye56.* Em termos estruturais, cada *cluster* pode ser organizado como destacado na Figura 20.

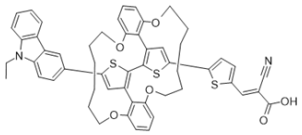
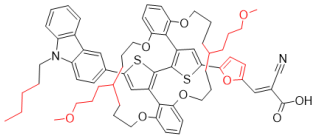
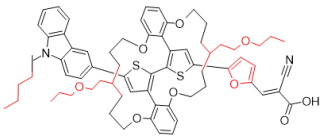
Figura 20 - Características estruturais para cada *cluster* gerado a partir de modificações no Dye56.



Dentre as 80 modificações apenas 13 delas apresentaram uma eficiência maior que 10%, porém 20 apresentaram um PCE predito menor que 8,91%. Do conjunto de estruturas modificadas, duas se destacam devido ao PCE predito exceder 11%. Essas duas modificações apresentam a substituição do anel tiofeno pelo furano, o aumento da cadeia carbônica ligada ao nitrogênio no grupo doador e a inserção de cadeias lineares no ciclo que envolve a ponte π.

A Tabela 27 apresenta a comparação entre a estrutura de partida e as modificações realizadas. Nota-se que as modificações proporcionaram um aumento e diminuição das energias do HOMO e LUMO, respectivamente. Provavelmente, essa modificação nas energias dos orbitais de fronteira permitirá uma maior compatibilidade energética entre a banda de condução do TiO_2 e o LUMO do corante, mas com um sutil afastamento entre o HOMO e o potencial de redução do par redox. Além disso, as modificações proporcionaram um deslocamento da absorção máxima calculada para o vermelho.

Tabela 27 – Comparação entre o Dye56 e as modificações estruturais com os maiores PCEs preditos.

Nome	Estruturas	PCE	HOMO (eV)	LUMO (eV)	λ_{max}^{Calc} (nm)
Dye56		9,20% **(E)**, 8,91% **(P)**	-8,029	-0,812	517,83
M-76		11,31%	-7,871	-1,278	561,72
M-80		11,04%	-7,902	-1,213	561,72

(E) – Experimental; **(P)** – Predito

Embora as estruturas M-76 e M-80 apresentem um PCE predito maior que 11%, outros estudos teóricos podem ser realizados para reforçar que essas modificações estruturais possivelmente proporcionarão DSSCs com PCE maiores que 9,2%. Por exemplo, estudos de transferência de carga intramolecular, injeção de carga no semicondutor e espontaneidade da formação do sistema corante + TiO_2 podem ser realizados com esse intuito.

5 CONCLUSÕES

Os DSSCs são dispositivos fotovoltaicos mais baratos, com forte potencial de serem desenvolvidos com uma alta capacidade de conversão. Isto tem motivado a pesquisa em busca de sensibilizadores que atinjam o objetivo, e a modelagem molecular tem se mostrado uma ferramenta importante. No presente trabalho, usou-se um conjunto de 126 corantes orgânicos baseados em carbazol e um total de 2393 variáveis para modelar a eficiência de conversão de uma DSSC.

A metodologia empregada para a construção dos modelos QSPR para o PCE mostrou-se pertinente, uma vez que foi possível obter um conjunto de 3 modelos (M-1, M-2 e M-4) com parâmetros de ajuste, predição e robustez que satisfazem os critérios de validação interna e externa comumente adotados.

O modelo mais preditivo, o M-4, indicou algumas características estruturais que os corantes devem apresentar para proporcionar maiores PCE. Desde características mais conhecidas como a absorção da região do visível até as posições menos favoráveis para a inserção de cadeias alquílicas para redução de interações π-π**.**

As modificações estruturais proporcionaram, no geral, a obtenção de estruturas que possivelmente proporcionem maiores eficiências de conversão para uma DSSCs que as utilizem. Com destaque para as modificações M-76 e M-80 que exibiram um PCE predito maior que 11%.

Desse modo, pode-se concluir que a abordagem empregada neste estudo QSPR resultou em um modelo com satisfatório poder de ajuste e predição, o qual foi empregado para proposição de estruturas de corantes baseados em carbazol com o intuito de obter DSSCs com maiores PCEs.

6 PERSPECTIVAS DO TRABALHO

- Modelar as outras propriedades fotovoltaicas (V_{OC} , J_{SC} e FF) e tentar interpretar o modelo mais preditivo para cada uma dessas propriedades;
- Melhorar o *script* utilizado para a obtenção dos modelos;
- Observar o efeito do solvente na otimização estrutural e, consequentemente, na obtenção de modelos;
- Estudar como as modificações alteraram o processo de transferência de carga intramolecular e a injeção de carga no semicondutor.

7 REFERÊNCIAS

[1] L. Dogaru, The main goals of the fourth industrial revolution. Renewable energy perspectives, in: Procedia Manufacturing, Elsevier B.V., 2020: pp. 397–401. https://doi.org/10.1016/j.promfg.2020.03.058.

[2] R. York, S.E. Bell, Energy transitions or additions?: Why a transition from fossil fuels requires more than the growth of renewable energy, Energy Research and Social Science. 51 (2019) 40–43. https://doi.org/10.1016/j.erss.2019.01.008.

[3] D. Gielen, F. Boshell, D. Saygin, M.D. Bazilian, N. Wagner, R. Gorini, The role of renewable energy in the global energy transformation, Energy Strategy Reviews. 24 (2019) 38–50. https://doi.org/10.1016/j.esr.2019.01.006.

[4] M. Grätzel, Solar energy conversion by dye-sensitized photovoltaic cells, Inorganic Chemistry. 44 (2005) 6841–6851. https://doi.org/10.1021/ic0508371.

[5] F. Arkan, M. Izadyar, Recent theoretical progress in the organic/metal-organic sensitizers as the free dyes, dye/TiO2 and dye/electrolyte systems; Structural modifications and solvent effects on their performance, Renewable and Sustainable Energy Reviews. 94 (2018) 609–655. https://doi.org/10.1016/j.rser.2018.06.054.

[6] Y. Liu, Y. Li, Y. Wu, G. Yang, L. Mazzarella, P. Procel-Moya, A.C. Tamboli, K. Weber, M. Boccard, O. Isabella, X. Yang, B. Sun, High-Efficiency Silicon Heterojunction Solar Cells: Materials, Devices and Applications, Materials Science and Engineering R: Reports. 142 (2020). https://doi.org/10.1016/j.mser.2020.100579.

[7] M. Petrović, V. Chellappan, S. Ramakrishna, Perovskites: Solar cells & engineering applications - materials and device developments, Solar Energy. 122 (2015) 678–699. https://doi.org/10.1016/j.solener.2015.09.041.

[8] D.M. Chapin, C.S. Fuller, G.L. Pearson, A new silicon p-n junction photocell for converting solar radiation into electrical power [3], Journal of Applied Physics. 25 (1954) 676–677. https://doi.org/10.1063/1.1721711.

[9] A. Blakers, N. Zin, K.R. McIntosh, K. Fong, High efficiency silicon solar cells, in: Energy Procedia, Elsevier Ltd, 2013: pp. 1–10. https://doi.org/10.1016/j.egypro.2013.05.033.

[10] E. Raphael, M.N. Silva, R. Szostak, M.A. Schiavon, A.F. Nogueira, CÉLULAS SOLARES DE PEROVSKITAS: UMA NOVA TECNOLOGIA EMERGENTE, Quimica Nova. 41 (2018) 61–74. https://doi.org/10.21577/0100-4042.20170127.

[11] A.B.F. Vitoreti, L.B. Corrêa, E. Raphael, A.O.T. Patrocínio, A.F. Nogueira, M.A. Schiavon, Células solares sensibilizadas por pontos quânticos, Química Nova. (2016). https://doi.org/10.21577/0100-4042.20160192.

[12] B. O'regan, M. Grätzel, A low-cost, high-efficiensy solar cell based on dye-sensitized colloidal TiO2 Films (GRATZEL, 1991), Nature. 353 (1991) 737–739.

[13] V. Venkatraman, B.K. Alsberg, A quantitative structure-property relationship study of the photovoltaic performance of phenothiazine dyes, Dyes and Pigments. 114 (2015) 69–77. https://doi.org/10.1016/j.dyepig.2014.10.026.

[14] V. Venkatraman, M. Foscato, V.R. Jensen, B.K. Alsberg, Evolutionary de novo design of phenothiazine derivatives for dye-sensitized solar cells, Journal of Materials Chemistry A. 3 (2015) 9851–9860. https://doi.org/10.1039/c5ta00625b.

[15] V. Venkatraman, P.O. Åstrand, B.K. Alsberg, Quantitative structure-property relationship modeling of Grätzel solar cell dyes, Journal of Computational Chemistry. 35 (2014) 214–226. https://doi.org/10.1002/jcc.23485.

[16] S.S. Soni, K.B. Fadadu, J. v. Vaghasiya, B.G. Solanki, K.K. Sonigara, A. Singh, D. Das, P.K. Iyer, Improved molecular architecture of D-π-A carbazole dyes: 9% PCE with a cobalt redox shuttle in dye sensitized solar cells, Journal of Materials Chemistry A. 3 (2015) 21664–21671. https://doi.org/10.1039/c5ta06548h.

[17] Z. Xu, X. Lu, Y. Li, S. Wei, Theoretical analysis on heteroleptic Cu(I)-based complexes for dye-sensitized solar cells: Effect of anchors on electronic structure, spectrum, excitation, and intramolecular and interfacial electron transfer, Molecules. 25 (2020). https://doi.org/10.3390/molecules25163681.

[18] G.G. Sonai, M.A.M. Jr, J.H.B. Nunes, J.D.M. Jr, A.F. Nogueira, Células solares sensibilizadas por corantes naturais: Um experimento introdutório sobre energia renovável para alunos de graduação, Quimica Nova. 38 (2015) 1357–1365. https://doi.org/10.5935/0100-4042.20150148.

[19] K. Sharma, V. Sharma, S.S. Sharma, Dye-Sensitized Solar Cells: Fundamentals and Current Status, Nanoscale Research Letters. 13 (2018). https://doi.org/10.1186/s11671-018-2760-6.

[20] B. Qi, J. Wang, Fill factor in organic solar cells, Physical Chemistry Chemical Physics. 15 (2013) 8972–8982. https://doi.org/10.1039/c3cp51383a.

[21] Y. Li, J. Liu, D. Liu, X. Li, Y. Xu, D-A-π-A based organic dyes for efficient DSSCs: A theoretical study on the role of π-spacer, Computational Materials Science. 161 (2019) 163–176. https://doi.org/10.1016/j.commatsci.2019.01.033.

[22] P. Naik, M.R. Elmorsy, R. Su, D.D. Babu, A. El-Shafei, A.V. Adhikari, New carbazole based metal-free organic dyes with D-π-A-π-A architecture for DSSCs: Synthesis, theoretical and cell performance studies, Solar Energy. 153 (2017) 600–610. https://doi.org/10.1016/J.SOLENER.2017.05.088.

[23] B. Nagarajan, S. Kushwaha, R. Elumalai, S. Mandal, K. Ramanujam, D. Raghavachari, Novel ethynyl-pyrene substituted phenothiazine based metal free organic dyes in DSSC with 12% conversion efficiency, Journal of Materials Chemistry A. 5 (2017) 10289–10300. https://doi.org/10.1039/c7ta01744h.

[24] X. Liao, H. Zhang, J. Huang, G. Wu, X. Yin, Y. Hong, (D−π−A)3−Type metal-free organic dye for dye-sensitized solar cells application, Dyes and Pigments. 158 (2018) 240–248. https://doi.org/10.1016/j.dyepig.2018.03.075.

[25] M.G. Murali, X. Wang, Q. Wang, S. Valiyaveettil, New banana shaped A–D–π–D–A type organic dyes containing two anchoring groups for high performance dye-sensitized solar cells, Dyes and Pigments. 134 (2016) 375–381. https://doi.org/10.1016/j.dyepig.2016.07.017.

[26] Ö. Birel, S. Nadeem, H. Duman, Porphyrin-Based Dye-Sensitized Solar Cells (DSSCs): a Review, Journal of Fluorescence. 27 (2017) 1075–1085. https://doi.org/10.1007/s10895-017-2041-2.

[27] L. Tian, X. Zhang, X. Xu, Z. Pang, X. Li, W. Wu, B. Liu, The planarization of side chain in carbazole sensitizer and its effect on optical, electrochemical, and interfacial charge transfer properties, Dyes and Pigments. 174 (2020). https://doi.org/10.1016/j.dyepig.2019.108036.

[28] D. Casanova, The role of the π linker in donor-π-acceptor organic dyes for high-performance sensitized solar cells, ChemPhysChem. 12 (2011) 2979–2988. https://doi.org/10.1002/cphc.201100520.

[29] L. Zhang, J.M. Cole, Anchoring groups for dye-sensitized solar cells, ACS Applied Materials and Interfaces. 7 (2015) 3427–3455. https://doi.org/10.1021/am507334m.

[30] S.F. Li, X.C. Yang, M. Cheng, J.H. Zhao, Y. Wang, L.C. Sun, Novel D–π–A type II organic sensitizers for dye sensitized solar cells, Tetrahedron Letters. 53 (2012) 3425–3428. https://doi.org/10.1016/J.TETLET.2012.04.049.

[31] M. Marszalek, S. Nagane, A. Ichake, R. Humphry-Baker, V. Paul, S.M. Zakeeruddin, M. Grätzel, Structural variations of D-π-A dyes influence on the photovoltaic performance of dye-sensitized solar cells, RSC Advances. 3 (2013) 7921–7927. https://doi.org/10.1039/c3ra22249g.

[32] Y. Ooyama, N. Yamaguchi, I. Imae, K. Komaguchi, J. Ohshita, Y. Harima, Dye-sensitized solar cells based on D–π–A fluorescent dyes with two pyridyl groups as an electron-withdrawing–injecting anchoring group, Chemical Communications. 49 (2013) 2548–2550. https://doi.org/10.1039/c3cc40498f.

[33] A. Yella, H.-W. Lee, H. Nok Tsao, C. Yi, A. Kumar Chandiran, M. Nazeeruddin, E. Wei-Guang Diau, C.-Y. Yeh, S.M. Zakeeruddin, M. Grätzel, Porphyrin-Sensitized Solar Cells with Cobalt (II/III)-Based Redox Electrolyte Exceed 12

Percent Efficiency, n.d. https://www.science.org/doi/abs/10.1126/science.1209688 (accessed June 7, 2022).

[34] J. Zhang, Y.H. Kan, H. bin Li, Y. Geng, Y. Wu, Z.M. Su, How to design proper π-spacer order of the D-π-A dyes for DSSCs? A density functional response, Dyes and Pigments. 95 (2012) 313–321. https://doi.org/10.1016/j.dyepig.2012.05.020.

[35] J. Gong, K. Sumathy, Q. Qiao, Z. Zhou, Review on dye-sensitized solar cells (DSSCs): Advanced techniques and research trends, Renewable and Sustainable Energy Reviews. 68 (2017) 234–246. https://doi.org/10.1016/j.rser.2016.09.097.

[36] S. Hwang, J.H. Lee, C. Park, H. Lee, C. Kim, C. Park, M.H. Lee, W. Lee, J. Park, K. Kim, N.G. Park, C. Kim, A highly efficient organic sensitizer for dye-sensitized solar cells, Chemical Communications. (2007) 4887–4889. https://doi.org/10.1039/b709859f.

[37] J. Liu, X. Yang, A. Islam, Y. Numata, S. Zhang, N.T. Salim, H. Chen, L. Han, Efficient metal-free sensitizers bearing circle chain embracing π-spacers for dye-sensitized solar cells, Journal of Materials Chemistry A. 1 (2013) 10889–10897. https://doi.org/10.1039/c3ta12368e.

[38] M.L. Han, Y.Z. Zhu, S. Liu, Q.L. Liu, D. Ye, B. Wang, J.Y. Zheng, The improved photovoltaic performance of phenothiazine-dithienopyrrole based dyes with auxiliary acceptors, Journal of Power Sources. 387 (2018) 117–125. https://doi.org/10.1016/j.jpowsour.2018.03.059.

[39] K. Kakiage, Y. Aoyama, T. Yano, K. Oya, J.I. Fujisawa, M. Hanaya, Highly-efficient dye-sensitized solar cells with collaborative sensitization by silyl-anchor and carboxy-anchor dyes, Chemical Communications. 51 (2015) 15894–15897. https://doi.org/10.1039/c5cc06759f.

[40] J. Yang, P. Ganesan, J. Teuscher, T. Moehl, Y.J. Kim, C. Yi, P. Comte, K. Pei, T.W. Holcombe, M.K. Nazeeruddin, J. Hua, S.M. Zakeeruddin, H. Tian, M. Grätzel, Influence of the donor size in D-π-A organic dyes for dye-sensitized solar cells, J Am Chem Soc. 136 (2014) 5722–5730. https://doi.org/10.1021/ja500280r.

[41] T. Delgado-Montiel, R. Soto-Rojo, J. Baldenebro-López, D. Glossman-Mitnik, Theoretical study of the effect of different π bridges including an azomethine group in triphenylamine-based dye for dye-sensitized solar cells, Molecules. 24 (2019). https://doi.org/10.3390/molecules24213897.

[42] G. Reginato, M. Calamante, L. Zani, A. Mordini, D. Franchi, Design and synthesis of organic sensitizers with enhanced anchoring stability in dye-sensitized solar cells, in: Pure and Applied Chemistry, Walter de Gruyter GmbH, 2018: pp. 363–376. https://doi.org/10.1515/pac-2017-0403.

[43] Z.S. Wang, Y. Cui, K. Hara, Y. Dan-Oh, C. Kasada, A. Shinpo, A high-light-harvesting-efficiency coumarin dye for stable dye-sensitized solar cells, Advanced Materials. 19 (2007) 1138–1141. https://doi.org/10.1002/adma.200601020.

[44] A. Dhar, N.S. Kumar, P.K. Paul, S. Roy, R.L. Vekariya, Influence of tagging thiophene bridge unit on optical and electrochemical properties of coumarin based dyes for DSSCs with theoretical insight, Organic Electronics. 53 (2018) 280–286. https://doi.org/10.1016/j.orgel.2017.12.007.

[45] Y. Ezhumalai, B. Lee, M.S. Fan, B. Harutyunyan, K. Prabakaran, C.P. Lee, S.H. Chang, J.S. Ni, S. Vegiraju, P. Priyanka, Y.W. Wu, C.W. Liu, S. Yau, J.T. Lin, C.G. Wu, M.J. Bedzyk, R.P.H. Chang, M.C. Chen, K.C. Ho, T.J. Marks, Metal-free branched alkyl tetrathienoacene (TTAR)-based sensitizers for high-performance dye-sensitized solar cells, Journal of Materials Chemistry A. 5 (2017) 12310–12321. https://doi.org/10.1039/c7ta01825h.

[46] D. Kim, J.K. Lee, S.O. Kang, J. Ko, Molecular engineering of organic dyes containing N-aryl carbazole moiety for solar cell, Tetrahedron. 63 (2007) 1913–1922. https://doi.org/10.1016/j.tet.2006.12.082.

[47] X. Hu, S. Cai, G. Tian, X. Li, J. Su, J. Li, Rigid triarylamine-based D-A-π-A structural organic sensitizers for solar cells: The significant enhancement of open-circuit photovoltage with a long alkyl group, RSC Advances. 3 (2013) 22544–22553. https://doi.org/10.1039/c3ra43057j.

[48] R. Todeschini, V. Consonni, Molecular Descriptors for Chemoinformatics, Volumes I & II, 2009.

[49] V.M. Alves, R.C. Braga, E.N. Muratov, C.H. Andrade, Cheminformatics: An introduction, Quimica Nova. 41 (2018) 202–212. https://doi.org/10.21577/0100-4042.20170145.

[50] A. Arroio, K.M. Honório, A.B.F. da Silva, Propriedades químico-quânticas empregadas em estudos das relações estrutura-atividade. Química Nova, 33, (2010) 694-699. doi.org/10.1590/S0100-40422010000300037.

[51] C. Selassie, R.P. Verma, History of Quantitative Structure–Activity Relationships, Burger's Medicinal Chemistry and Drug Discovery. (2010) 1–96. https://doi.org/10.1002/0471266949.BMC001.PUB2.

[52] E. Directorate, ENV/JM/MONO(2007)2 2 OECD Environment Health and Safety Publications Series on Testing and Assessment No. 69 Guidance Document On The Validation Of (Quantitative) Structure-Activity Relationship [(Q)SAR] Models, 2007.

[53] V. Consonni, D. Ballabio, R. Todeschini, Comments on the definition of the Q2 parameter for QSAR validation, Journal of Chemical Information and Modeling. 49 (2009) 1669–1678. https://doi.org/10.1021/ci900115y.

[54] K. Roy, S. Kar, R.N. Das, Statistical Methods in QSAR/QSPR, in: 2015: pp. 37–59. https://doi.org/10.1007/978-3-319-17281-1_2.

[55] P. Kawczak, L. Bober, T. Bączek, Activity evaluation of some psychoactive drugs with the application of QSAR/QSPR modeling methods, Medicinal Chemistry Research. 27 (2018) 2279–2286. https://doi.org/10.1007/s00044-018-2234-5.

[56] J. Xu, H. Zhang, L. Wang, G. Liang, L. Wang, X. Shen, W. Xu, QSPR study of absorption maxima of organic dyes for dye-sensitized solar cells based on 3D descriptors, Spectrochimica Acta - Part A: Molecular and Biomolecular Spectroscopy. 76 (2010) 239–247. https://doi.org/10.1016/j.saa.2010.03.027.

[57] M.M.C. Ferreira, Quimiometria: conceitos, métodos e aplicações, Editora da Unicamp, 2015. https://doi.org/10.7476/9788526814714.

[58] H. Abdi, L.J. Williams, Principal component analysis, Wiley Interdisciplinary Reviews: Computational Statistics. 2 (2010) 433–459. https://doi.org/10.1002/wics.101.

[59] M. Otto, Chemometrics : statistics and computer application in analytical chemistry, (2017).

[60] R. Gelbard, O. Goldman, I. Spiegler, Investigating diversity of clustering methods: An empirical comparison, Data and Knowledge Engineering. 63 (2007) 155–166. https://doi.org/10.1016/j.datak.2007.01.002.

[61] P. Gramatica, Principles of QSAR Modeling, International Journal of Quantitative Structure-Property Relationships. 5 (2020) 61–97. https://doi.org/10.4018/ijqspr.20200701.oa1.

[62] N.Richard. Draper, H. Smith, Applied regression analysis, (1998) 706.

[63] G. Heinze, C. Wallisch, D. Dunkler, Variable selection – A review and recommendations for the practicing statistician, Biometrical Journal. 60 (2018) 431–449. https://doi.org/10.1002/bimj.201700067.

[64] S. Forrest, Genetic Algorithms: Principles of Natural Selection Applied to Computation, Science (1979). 261 (1993) 872–878. https://doi.org/10.1126/SCIENCE.8346439.

[65] L. Eriksson, J. Jaworska, A.P. Worth, M.T.D. Cronin, R.M. McDowell, P. Gramatica, Methods for reliability and uncertainty assessment and for applicability evaluations of classification- and regression-based QSARs, Environmental Health Perspectives. 111 (2003) 1361–1375. https://doi.org/10.1289/ehp.5758.

[66] H. Rajak, S. Sivadasan, Validation of QSAR Models-Strategies and Importance, 2011. https://www.researchgate.net/publication/284566093.

[67] A. Golbraikh, A. Tropsha, Beware of q2!, Journal of Molecular Graphics and Modelling. 20 (2002) 269–276. https://doi.org/10.1016/S1093-3263(01)00123-1.

[68] P.P. Roy, K. Roy, On some aspects of variable selection for partial least squares regression models, QSAR and Combinatorial Science. 27 (2008) 302–313. https://doi.org/10.1002/qsar.200710043.

[69] P.P. Roy, S. Paul, I. Mitra, K. Roy, On two novel parameters for validation of predictive QSAR models, Molecules. 14 (2009) 1660–1701. https://doi.org/10.3390/molecules14051660.

[70] K. Roy, I. Mitra, S. Kar, P.K. Ojha, R.N. Das, H. Kabir, Comparative studies on some metrics for external validation of QSPR models, in: Journal of Chemical Information and Modeling, American Chemical Society, 2012: pp. 396–408. https://doi.org/10.1021/ci200520g.

[71] A. Tropsha, Best practices for QSAR model development, validation, and exploitation, Molecular Informatics. 29 (2010) 476–488. https://doi.org/10.1002/minf.201000061.

[72] V. Venkatraman, L.K. Chellappan, An open access data set highlighting aggregation of dyes on metal oxides, Data (Basel). 5 (2020). https://doi.org/10.3390/data5020045.

[73] Stewart Computational Chemistry - MOPAC Home Page, (n.d.). http://openmopac.net/ (accessed June 8, 2022).

[74] C.W. Yap, PaDEL-descriptor: An open source software to calculate molecular descriptors and fingerprints, Journal of Computational Chemistry. 32 (2011) 1466–1474. https://doi.org/10.1002/jcc.21707.

[75] R: The R Project for Statistical Computing, (n.d.). https://www.r-project.org/ (accessed June 8, 2022).

[76] RStudio | Open source & professional software for data science teams - RStudio, (n.d.). https://www.rstudio.com/ (accessed June 8, 2022).

[77] CRAN - Package FactoMineR, (n.d.). https://cran.r-project.org/web/packages/FactoMineR/index.html (accessed June 8, 2022).

[78] CRAN - Package gaselect, (n.d.). https://cran.r-project.org/web/packages/gaselect/index.html (accessed June 8, 2022).

[79] CRAN - Package leaps, (n.d.). https://cran.r-project.org/web/packages/leaps/index.html (accessed June 8, 2022).

[80] CRAN - Package Metrics, (n.d.). https://cran.r-project.org/web/packages/Metrics/index.html (accessed June 8, 2022).

[81] A. Tropsha, P. Gramatica, V.K. Gombar, The importance of being earnest: Validation is the absolute essential for successful application and interpretation of QSPR models, in: QSAR and Combinatorial Science, Wiley-VCH Verlag, 2003: pp. 69–77. https://doi.org/10.1002/qsar.200390007.

[82] D.T. Stanton, P.C. Jurs, Development and Use of Charged Partial Surface Area Structural Descriptors in Computer-Assisted Quantitative Structure-Property Relationship Studies, Analytical Chemistry. 62 (1990) 2323–2329. https://doi.org/10.1021/AC00220A013/ASSET/AC00220A013.FP.PNG_V03.

[83] B. Liu, R. Wang, W. Mi, X. Li, H. Yu, Novel branched coumarin dyes for dye-sensitized solar cells: significant improvement in photovoltaic performance by simple structure modification, Journal of Materials Chemistry. 22 (2012) 15379–15387. https://doi.org/10.1039/C2JM32333H.

[84] L.F. Lai, C.L. Ho, Y.C. Chen, W.J. Wu, F.R. Dai, C.H. Chui, S.P. Huang, K.P. Guo, J.T.S. Lin, H. Tian, S.H. Yang, W.Y. Wong, New bithiazole-functionalized organic photosensitizers for dye-sensitized solar cells, Dyes and Pigments. 96 (2013) 516–524. https://doi.org/10.1016/J.DYEPIG.2012.10.002.

[85] L.H. Hall, L.B. Kier, Electrotopological State Indices for Atom Types: A Novel Combination of Electronic, Topological, and Valence State Information, Journal of Chemical Information and Computer Sciences. 35 (1995) 1039–1045. https://doi.org/10.1021/CI00028A014/ASSET/CI00028A014.FP.PNG_V03.

[86] L.H. Hall, B. Mohney, L.B. Kier, The Electrotopological State: Structure Information at the Atomic Level for Molecular Graphs, Journal of Chemical Information and Computer Sciences. 31 (1991) 76–82. https://doi.org/10.1021/CI00001A012/ASSET/CI00001A012.FP.PNG_V03.

[87] F.R. Burden, A Chemically Intuitive Molecular Index Based on the Eigenvalues of a Modified Adjacency Matrix, Quantitative Structure-Activity Relationships. 16 (1997) 309–314. https://doi.org/10.1002/QSAR.19970160406.

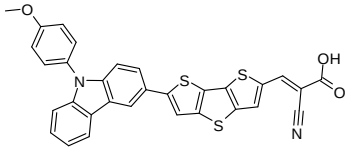
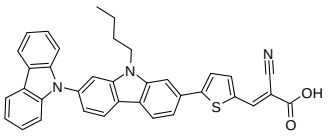
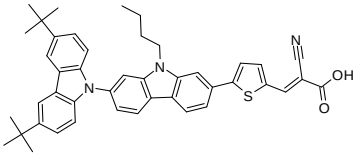
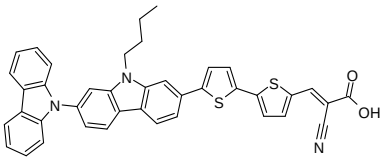
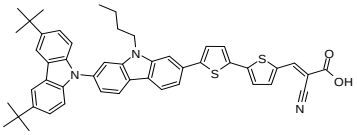
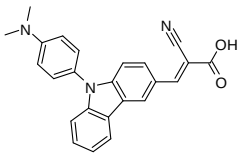
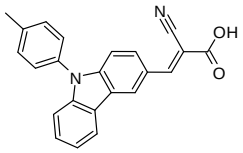
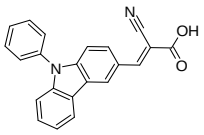
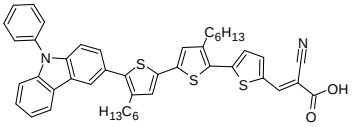
[88] F.R. Burden, Molecular identification number for substructure searches, Journal of Chemical Information and Computer Sciences. 29 (1989) 225–227. https://doi.org/10.1021/CI00063A011/ASSET/CI00063A011.FP.PNG_V03.

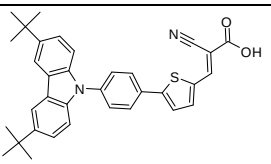
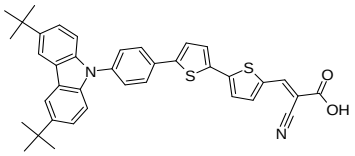
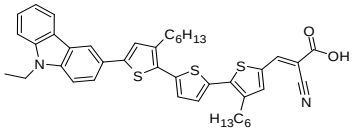
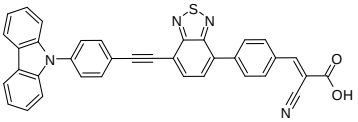
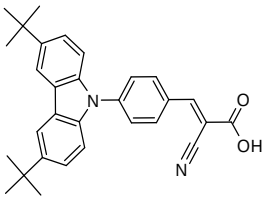
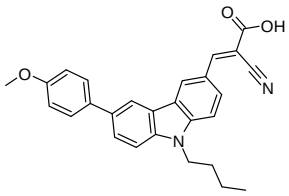
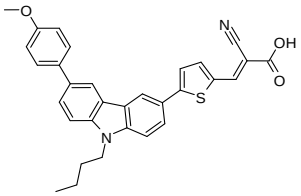
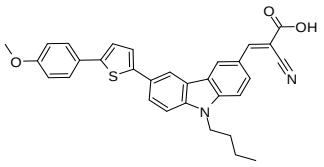
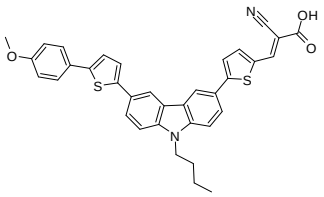
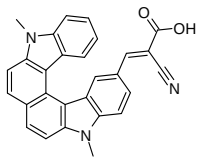
[89] R.S. Pearlman, K.M. Smith, Novel software tools for chemical diversity, Perspectives in Drug Discovery and Design 1998 9:0. 9 (1998) 339–353. https://doi.org/10.1023/A:1027232610247.

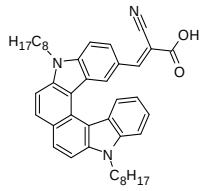
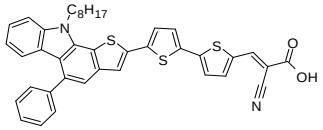
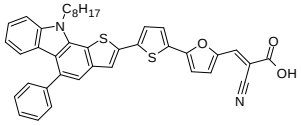
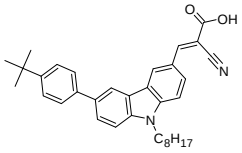
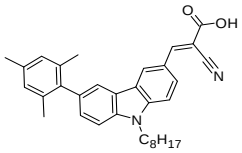
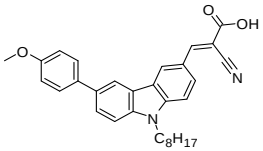
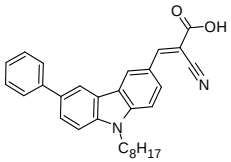
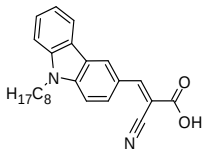
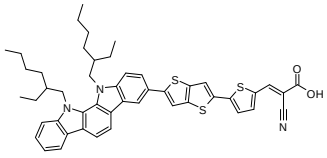
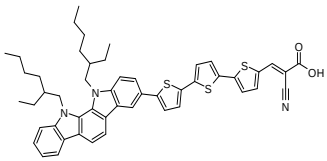
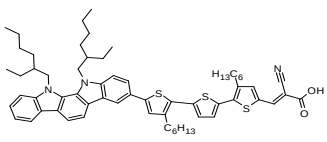
8 APÊNDICES

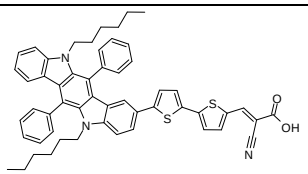
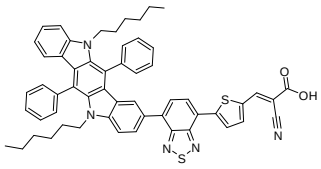
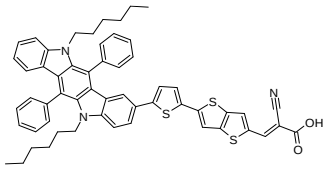
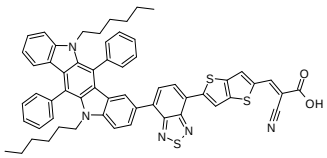
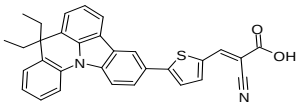
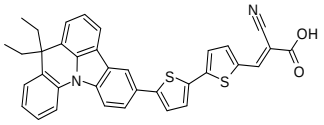
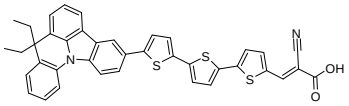
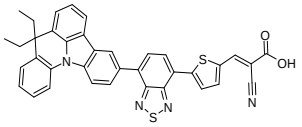
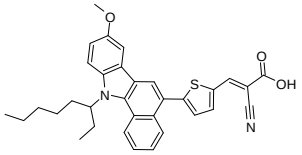
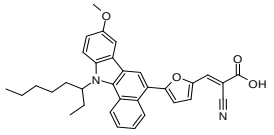
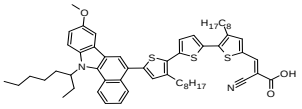
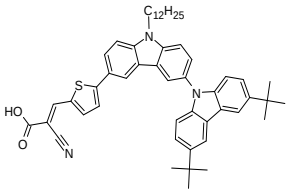
8.1 Estruturas dos Corantes e Performances Fotovoltaicas

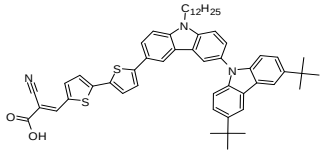
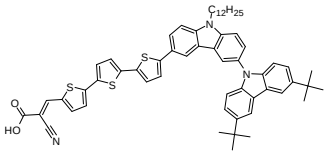
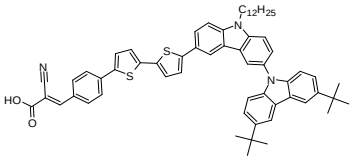
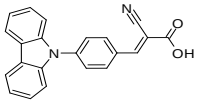
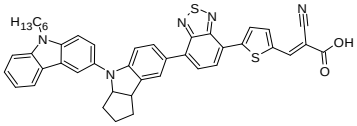
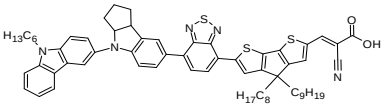
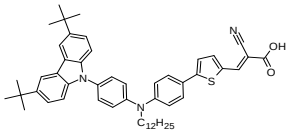
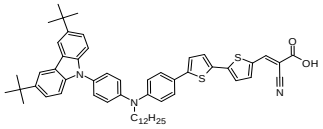
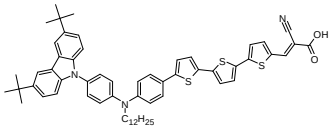
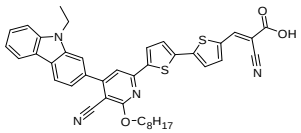
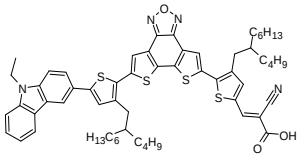
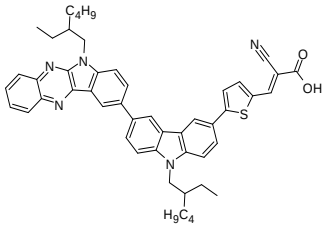
Ac – Acetonitrila; **Et –** Etanol; **Di –** Diclorometano; **DMF -** Dimetilformida**; Met** – Metanol**;**
Tr – Triclorometano; **THF** – Tetrahidrofurano;

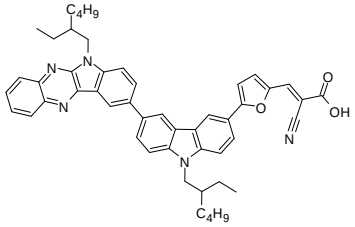
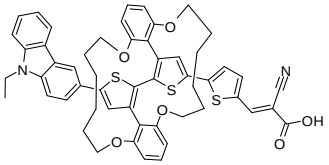
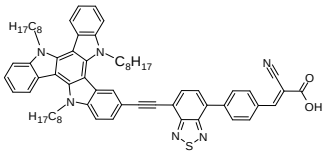
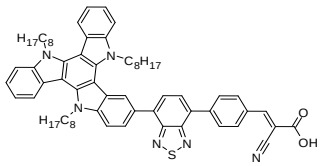
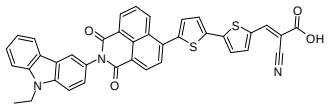
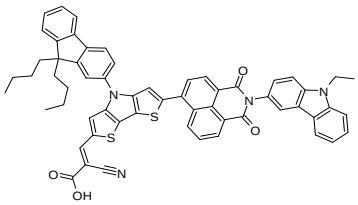
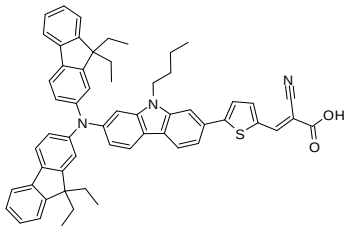
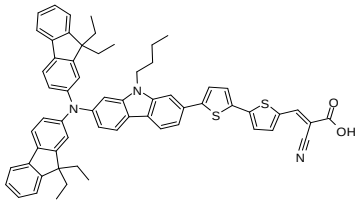
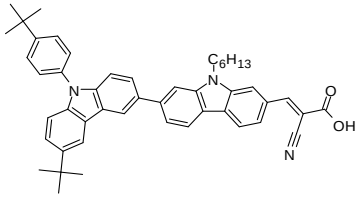
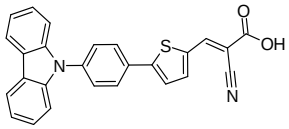
Dye	Estrutura do Corante	%PCE	J_{sc} ($mA.cm^{-2}$)	V_{oc}(V)	FF	Sol.	doi
1.		5,64	13,45	0,618	0,680	Et	10.1016/j.tet.2013.02.058
2.		4,22	9,98	0,635	0,660	Di	10.1021/am404948w
3.		4,95	11,22	0,666	0,660	Di	10.1021/am404948w
4.		6,04	15,78	0,601	0,640	Di	10.1021/am404948w
5.		5,48	14,00	0,612	0,640	Di	10.1021/am404948w
6.		1,21	3,09	0,640	0,610	Ac	10.1108/PRT-09-2014-0077
7.		2,82	7,22	0,640	0,610	Ac	10.1108/PRT-09-2014-0077
8.		3,69	9,29	0,630	0,630	Ac	10.1108/PRT-09-2014-0077
9.		5,92	13,60	0,740	0,589	Ac	10.1016/j.jpowsour.2020.227776

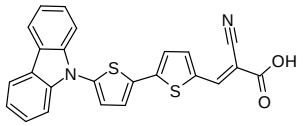
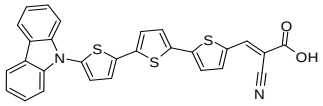
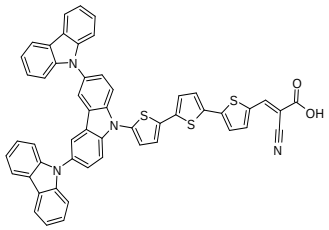
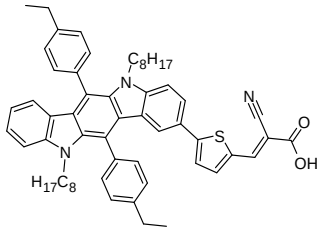
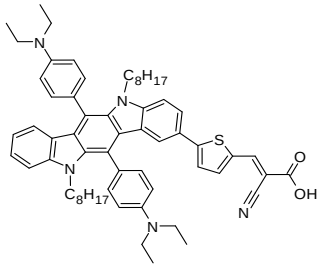
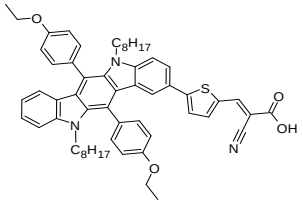
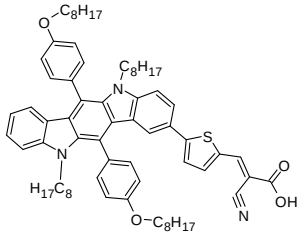
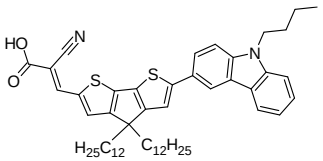
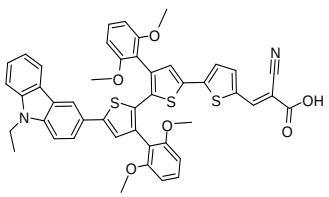
10.		2,39	8,47	0,470	0,600	Tr	doi.org/10.1021/jp1055842
11.		2,48	10,9	0,400	0,570	Tr	doi.org/10.1021/jp1055842
12.		7,44	15,73	0,701	0,670	Tr	10.1016/j.electacta.2018.08.068
13.		3,50	7,95	0,640	0,690	Tr	[7]10.1016/j.dyepig.2016.08.013
14.		2,68	6,82	0,577	0,681	Ac	10.1246/cl.2010.864
15.		1,87	3,69	0,720	0,700	Di	10.1002/ejoc.201600353
16.		4,54	8,47	0,760	0,700	Di	10.1002/ejoc.201600353
17.		2,52	4,70	0,740	0,730	Di	10.1002/ejoc.201600353
18.		4,57	8,59	0,750	0,710	Di	10.1002/ejoc.201600353
19.		2,49	7,46	0,560	0,600	Tr	10.1016/j.solmat.2009.11.014

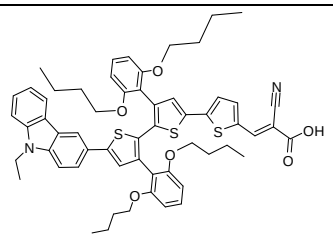
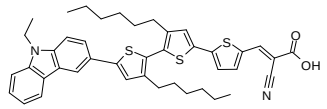
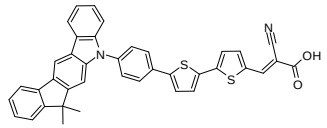
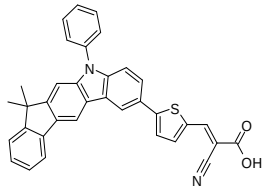
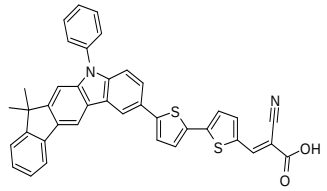
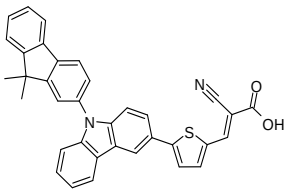
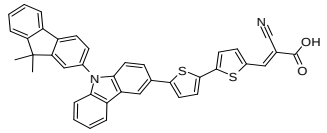
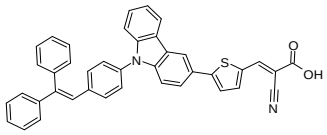
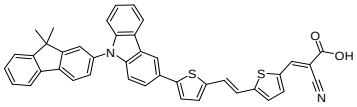
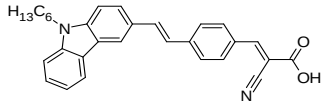
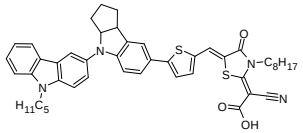
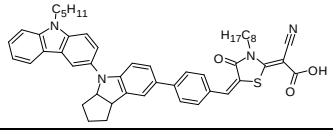
20.		3,18	8,40	0,660	0,570	Tr	10.1016/j.sol-mat.2009.11.014
21.		6,6	12,40	0,700	0,759	Tr	10.1016/j.tet.2014.01.001
22.		6,73	12,49	0,710	0,756	Tr	10.1016/j.tet.2014.01.001
23.		2,17	4,15	0,730	0,718	Tr	10.1039/C6RA01185C
24.		0,98	2,06	0,634	0,750	Tr	10.1039/C6RA01185C
25.		2,69	4,96	0,730	0,743	Tr	10.1039/C6RA01185C
26.		0,98	2,24	0,574	0,765	Tr	10.1039/C6RA01185C
27.		1,11	2,63	0,577	0,731	Tr	10.1039/C6RA01185C
28.		5,78	10,73	0,731	0,737	Tr	10.1016/j.dyepig.2019.01.033
29.		5,23	9,81	0,680	0,784	Tr	10.1016/j.dyepig.2019.01.033
30.		5,97	10,95	0,754	0,723	Tr	10.1016/j.dyepig.2019.01.033

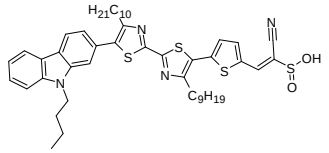
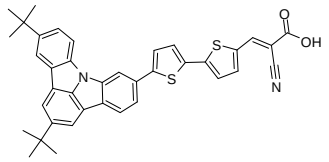
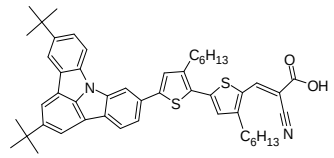
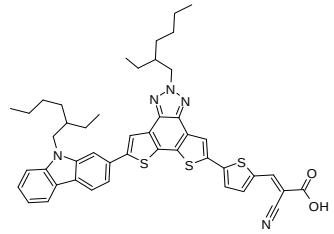
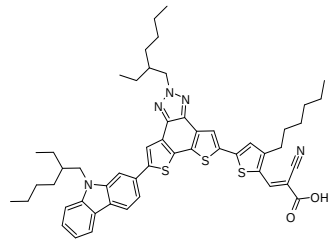
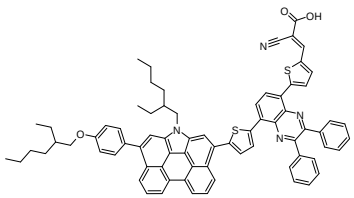
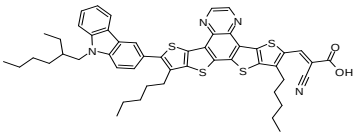
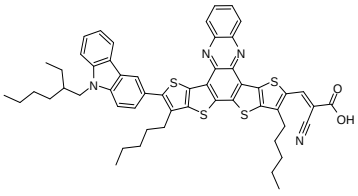
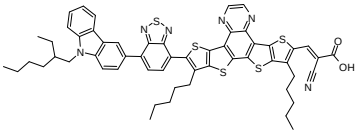
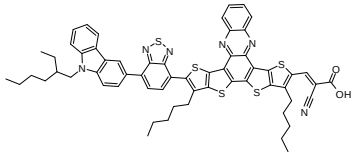
31.		6,09	11,95	0,768	0,660	THF	10.1039/C3TA 11748K
32.		5,55	11,57	0,707	0,680	THF	10.1039/C3TA 11748K
33.		4,11	9,40	0,644	0,680	THF	10.1039/C3TA 11748K
34.		6,40	13,96	0,674	0,680	THF	10.1039/C3TA 11748K
35.		5,43	10,4	0,740	0,702	THF	10.1039/C3TA 01657A
36.		6,50	13,7	0,690	0,691	THF	10.1039/C3TA 01657A
37.		2,96	7,57	0,570	0,689	THF	10.1039/C3TA 01657A
38.		4,61	10,0	0,640	0,722	THF	10.1039/C3TA 01657A
39.		6,01	12,4	0,729	0,660	Di	10.1021/am50 8400a
40.		6,93	13,8	0,757	0,660	Di	10.1021/am50 8400a
41.		7,54	14,8	0,744	0,680	Di	10.1021/am50 8400a
42.		3,64	7,35	0,740	0,670	Di	10.1002/ejoc.2 01300373

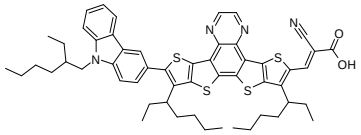
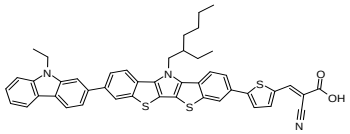
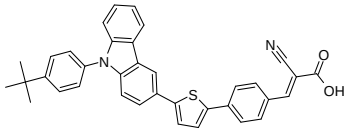
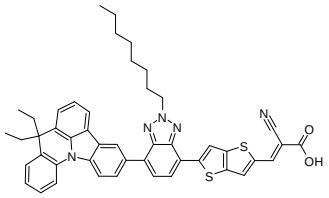
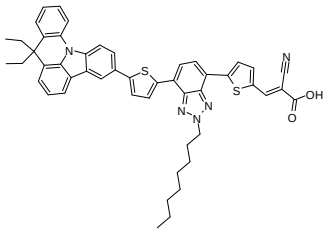
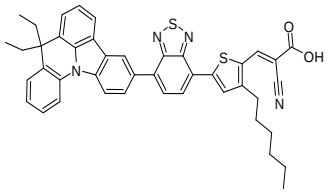
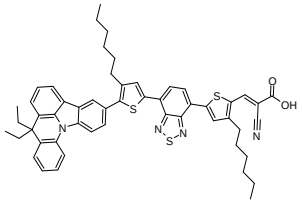
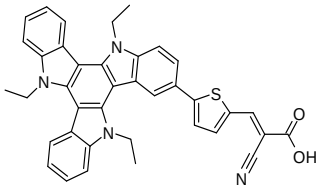
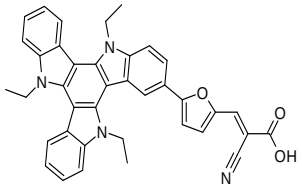
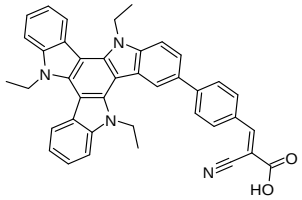
43.		4,80	9,70	0,730	0,680	Di	10.1002/ejoc.201300373
44.		5,69	11,31	0,710	0,710	Di	10.1002/ejoc.201300373
45.		4,62	9,94	0,700	0,670	Di	10.1002/ejoc.201300373
46.		1,77	4,25	0,586	0,710	Di	10.1016/j.dyepig.2012.03.028
47.		5,13	12,21	0,634	0,660	Di	10.1021/acsami.5b08888
48.		7,69	16,87	0,696	0,660	Di	10.1021/acsami.5b08888
49.		3,52	7,19	0,730	0,670	Di	10.1021/jp304489t
50.		4,10	8,88	0,700	0,660	Di	10.1021/jp304489t
51.		5,12	10,89	0,700	0,670	Di	10.1021/jp304489t
52.		3,34	9,01	0,550	0,670	Di	10.1016/j.solener.2018.09.073
53.		5,98	12,43	0,680	0,720	THF	10.1021/am5067145
54.		6,48	14,8	0,701	0,630	Di	10.1016/j.jpowsour.2015.01.148

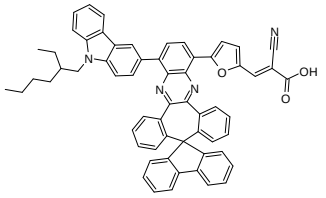
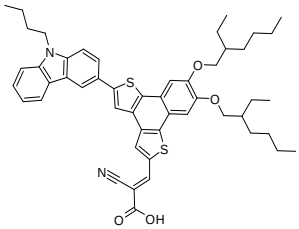
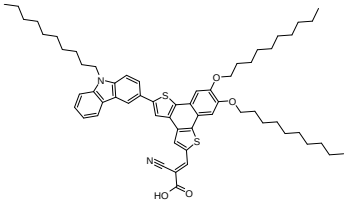
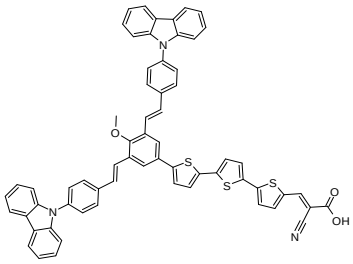
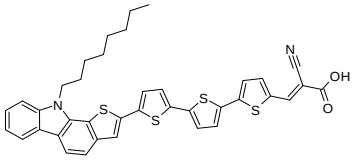
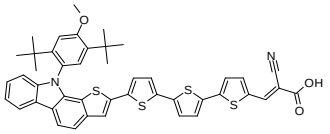
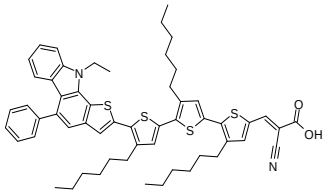
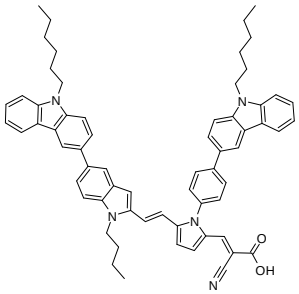
55.		7,03	14,1	0,742	0,670	Di	10.1016/j.jpowsour.2015.01.148
56.		9,20	16,34	0,747	0,754	Di	10.1039/C3TA12368E
57.		7,15	16,45	0,707	0,615	Di	10.1039/C7NJ04629D
58.		7,26	15,89	0,743	0,615	Di	10.1039/C7NJ04629D
59.		0,57	3,81	0,380	0,490	Di	10.1007/s10854-018-9750-4
60.		0,92	3,60	0,519	0,490	Di	10.1007/s10854-018-9750-4
61.		6,44	15,60	0,666	0,620	Di	10.1016/j.dyepig.2015.07.034
62.		4,77	12,50	0,630	0,610	Di	10.1016/j.dyepig.2015.07.034
63.		4,38	6,59	0,930	0,720	Di	10.1016/j.dyepig.2015.09.004
64.		2,74	6,17	0,601	0,740	Di	10.1016/j.dyepig.2013.09.025

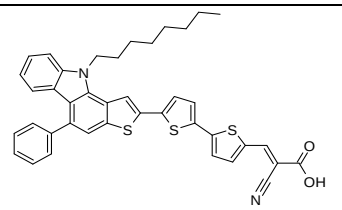
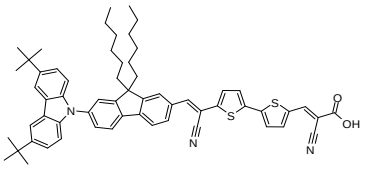
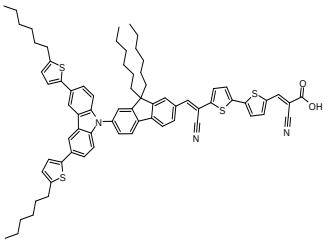
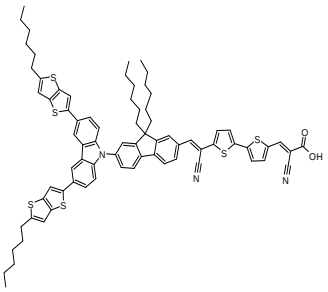
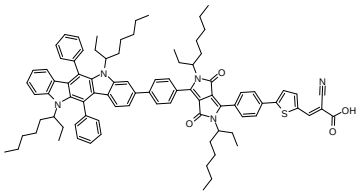
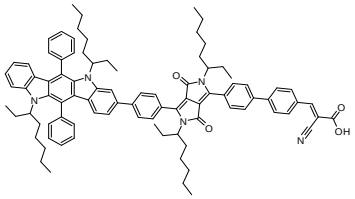
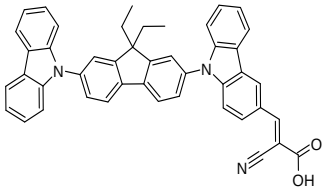
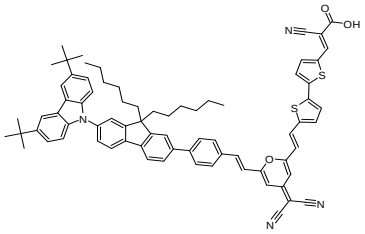
65.		2,94	6,70	0,589	0,740	Di	10.1016/j.dye-pig.2013.09.025
66.		4,30	9,98	0,586	0,740	Di	10.1016/j.dye-pig.2013.09.025
67.		4,86	10,65	0,643	0,710	Di	10.1016/j.dye-pig.2013.09.025
68.		6,25	12,6	0,729	0,680	Di	10.1016/j.jpowsour.2016.04.043
69.		8,09	15,2	0,745	0,710	Di	10.1016/j.jpowsour.2016.04.043
70.		6,98	13,9	0,738	0,680	Di	10.1016/j.jpowsour.2016.04.043
71.		7,58	14,1	0,757	0,710	Di	10.1016/j.jpowsour.2016.04.043
72.		7,5	13,1	0,770	0,730	Di	10.1039/C3RA22249G
73.		7,01	12,94	0,738	0,734	Di	10.1039/C9TC01520E

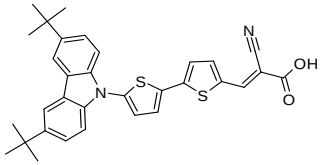
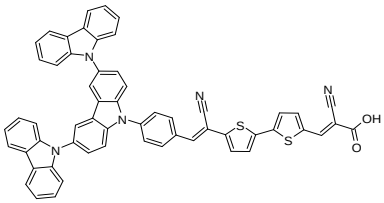
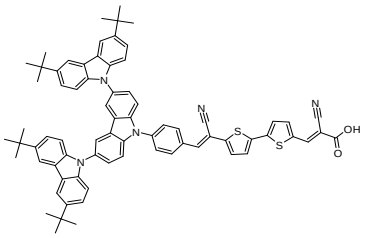
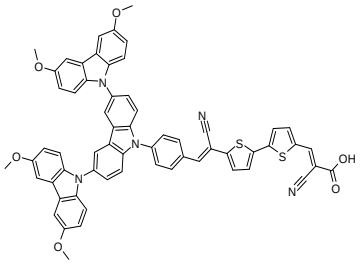
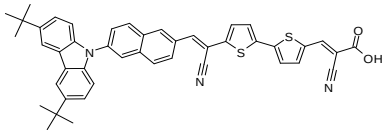
74.		8,01	13,05	0,801	0,766	Di	10.1039/C9TC01520E
75.		5,06	9,96	0,715	0,711	Di	10.1039/C9TC01520E
76.		4,23	8,55	0,640	0,770	DMF	10.1039/C7PP00350A
77.		5,97	12,34	0,710	0,680	DMF	10.1039/C7PP00350A
78.		5,34	11,68	0,670	0,680	DMF	10.1039/C7PP00350A
79.		5,02	9,83	0,740	0,700	Et	10.1016/j.tet.2006.12.082
80.		5,15	11,50	0,680	0,660	Et	10.1016/j.tet.2006.12.082
81.		3,87	7,85	0,690	0,710	Et	10.1016/j.tet.2006.12.082
82.		3,76	9,11	0,610	0,680	Et	10.1016/j.tet.2006.12.082
83.		7,1	17,2	0,645	0,650	Et	10.1039/C5TA06548H
84.		8,48	17,49	0,713	0,681	Met	10.1016/j.dyepig.2018.03.072
85.		4,69	8,86	0,727	0,728	Met	10.1016/j.dyepig.2018.03.072

86.		4,65	9,61	0,695	0,700	THF	10.1016/j.dyepig.2012.10.002
87.		3,96	10,41	0,530	0,720	THF	10.1039/C5RA02720A
88.		2,85	7,01	0,550	0,740	THF	10.1039/C5RA02720A
89.		7,52	14,46	0,725	0,720	THF	10.1039/C6TA02275H
90.		8,51	15,76	0,744	0,730	THF	10.1039/C6TA02275H
91.		7,58	14,89	0,728	0,700	THF	10.1016/j.dyepig.2016.12.013
92.		6,48	12,92	0,661	0,757	THF	10.1016/j.dye-pig.2018.06.010
93.		6,33	12,47	0,695	0,731	THF	10.1016/j.dye-pig.2018.06.010
94.		7,77	15,63	0,691	0,719	THF	10.1016/j.dye-pig.2018.06.010
95.		5,23	13,76	0,691	0,748	THF	10.1016/j.dye-pig.2018.06.010

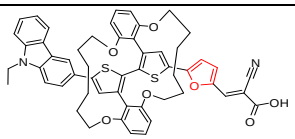
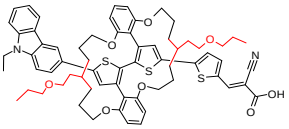
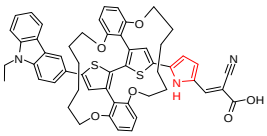
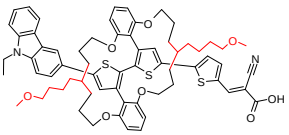
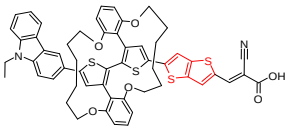
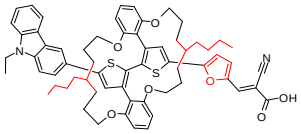
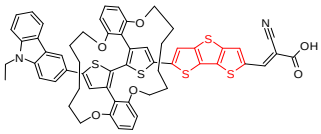
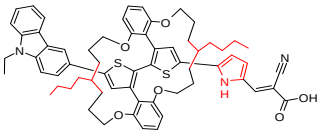
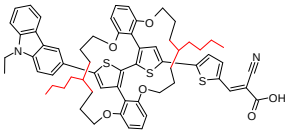
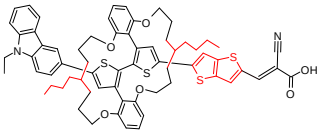
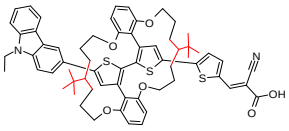
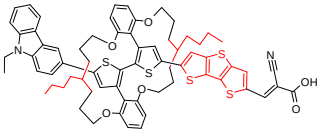
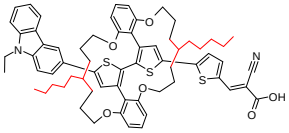
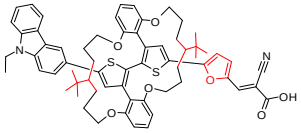
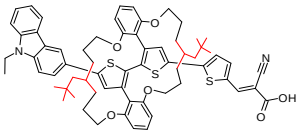
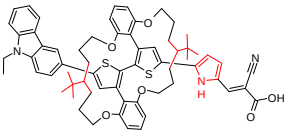
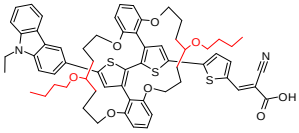
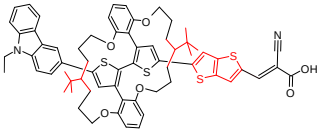
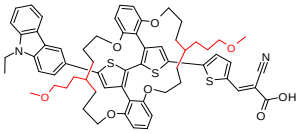
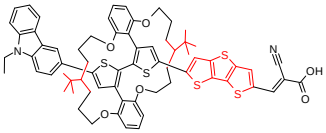
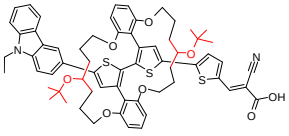
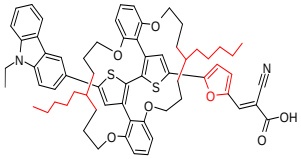
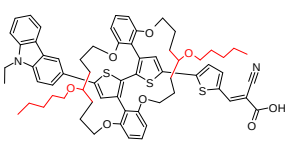
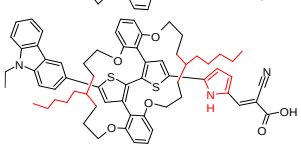
96.		5,23	11,08	0,641	0,735	THF	10.1016/j.dye-pig.2018.06.010
97.		5,30	12,75	0,645	0,640	THF	10.1016/j.dyepig.2018.10.004
98.		5,65	9,54	0,808	0,730	THF	10.1039/C3RA43057J
99.		6,23	10,56	0,829	0,710	THF	10.1039/C3RA43057J
100.		7,15	13,45	0,757	0,700	THF	10.1039/C3RA43057J
101.		5,20	9,71	0,712	0,750	THF	10.1039/C3RA43057J
102.		5,82	11,82	0,682	0,720	THF	10.1039/C3RA43057J
103.		6,10	14,7	0,670	0,620	THF	10.1021/ol402931u
104.		5,50	13,6	0,654	0,620	THF	10.1021/ol402931u
105.		5,11	11,6	0,689	0,640	THF	10.1021/ol402931u

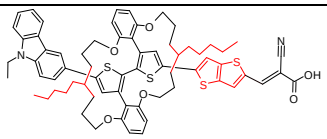
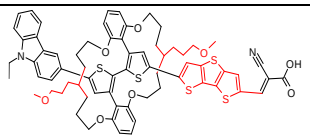
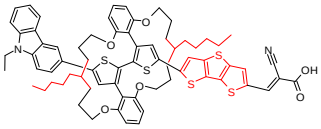
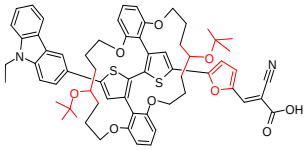
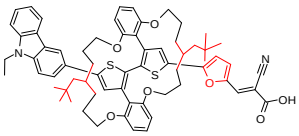
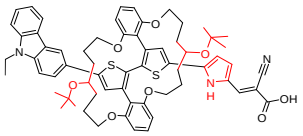
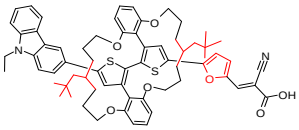
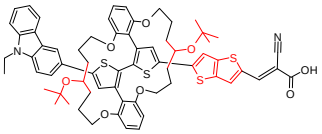
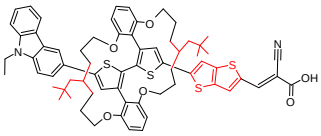
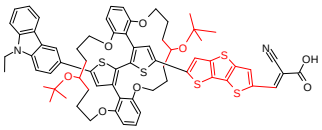
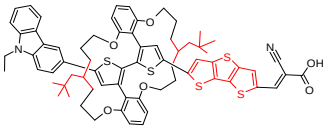
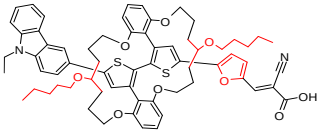
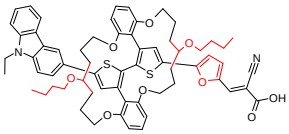
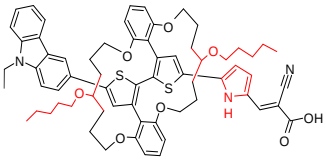
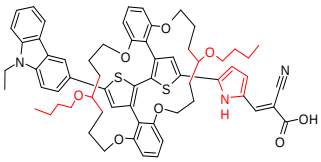
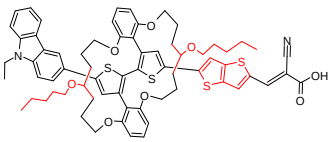
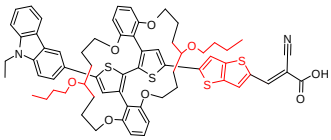
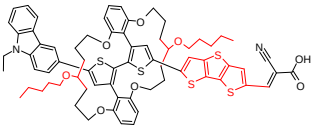
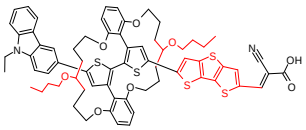
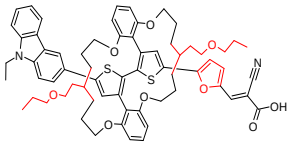
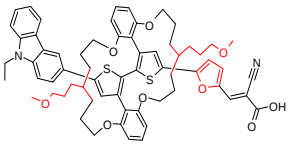
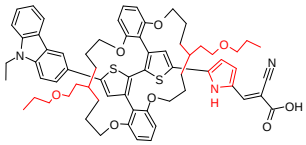
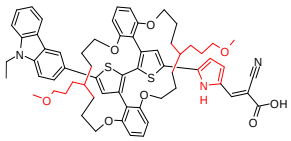
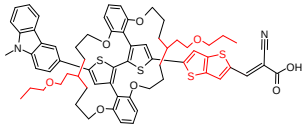
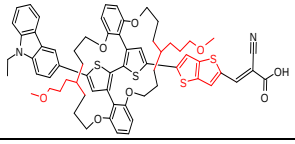
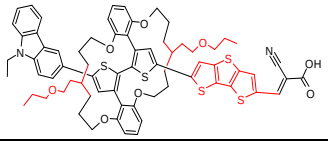
106.		4,87	11,00	0,710	0,627	Di	10.1002/gch2.201900034
107.		4,49	10,6	0,657	0,645	THF	10.1039/C3TA12901B
108.		4,60	10,8	0,663	0,640	THF	10.1039/C3TA12901B
109.		3,03	8,19	0,550	0,693	THF	10.1002/asia.201402654
110.		5,9	13,2	0,630	0,700	Tr	10.1039/C4QO00285G
111.		6,5	13,1	0,680	0,730	Tr	10.1039/C4QO00285G
112.		7,0	13,9	0,740	0,680	Tr	10.1039/C4QO00285G
113.		4,31	10,25	0,690	0,610	Et	10.1021/jp906334w

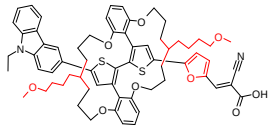
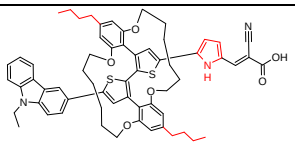
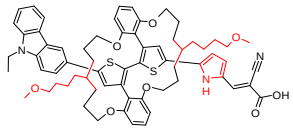
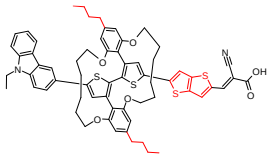
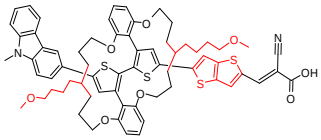
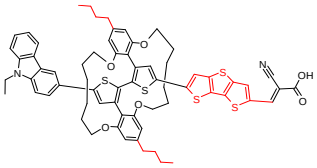
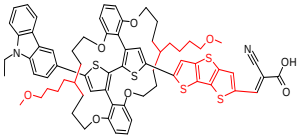
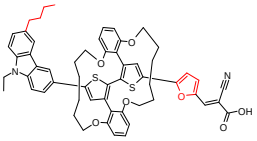
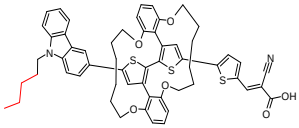
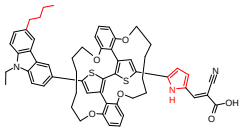
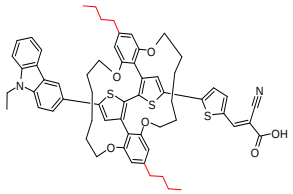
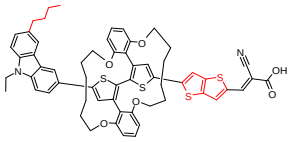
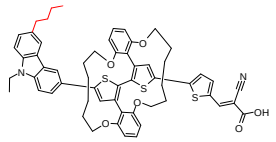
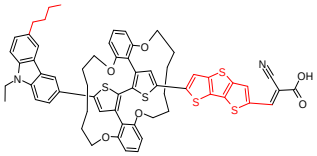
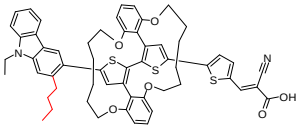
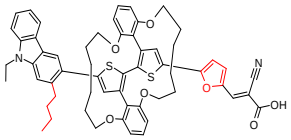
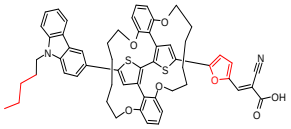
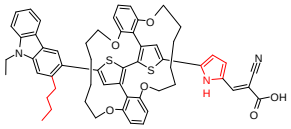
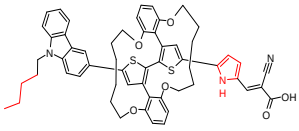
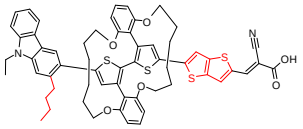
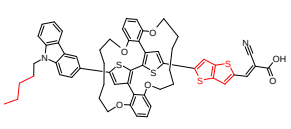
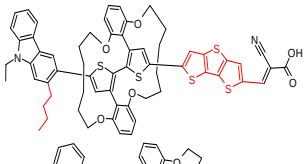
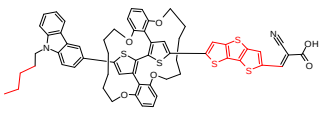
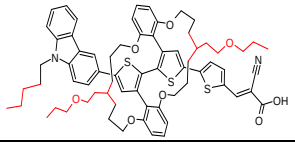
114.		5,96	12,30	0,691	0,701	Tr	10.1016/j.tet.2015.04.018
115.		5,2	12,4	0,627	0,670	Di	10.1016/j.dyepig.2015.02.020
116.		6,5	14,4	0,620	0,720	Di	10.1016/j.dyepig.2015.02.020
117.		6,5	14,0	0,632	0,740	Di	10.1016/j.dyepig.2015.02.020
118.		6,95	14,28	0,680	0,720	Di	10.1039/C4TA05162A
119.		6,67	13,63	0,691	0,710	Di	10.1039/C4TA05162A
120.		2,30	5,46	0,605	0,700	Di	10.1021/jo200501b
121.		3,19	7,14	0,620	0,720	Di	10.1016/j.tetlet.2014.04.037

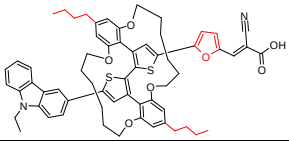
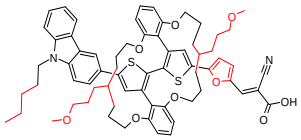
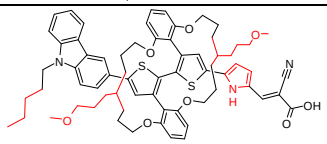
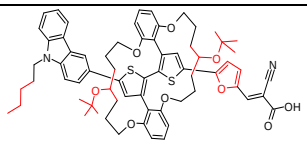
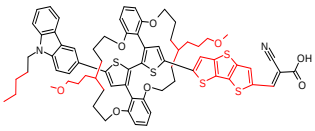
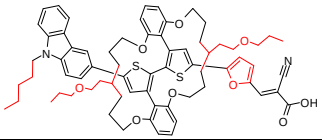
122.		5,10	9,89	0,720	0,730	Di	10.1021/am50 0947k
123.		4,90	11,70	0,580	0,724	Di	10.1002/cssc.2 01200975
124.		5,80	13,40	0,590	0,723	Di	10.1002/cssc.2 01200975
125.		5,80	14,30	0,560	0,723	Di	10.1002/cssc.2 01200975
126.		5,60	13,80	0,580	0,693	Di	10.1002/cssc.2 01200975

8.2 Modificações no corante *Dye56*

Nº	Estrutura	PCE_{Pred}	Cluster	Nº	Estrutura	PCE_{Pred}	Cluster
M-1		9.49	1	M-13		10.15	2
M-2		8.86	1	M-14		9.77	3
M-3		8.61	4	M-15		9.85	2
M-4		8.51	3	M-16		8.99	2
M-5		9.03	2	M-17		8.83	3
M-6		6.90	2	M-18		9.51	3
M-7		8.98	2	M-19		7.90	2
M-8		9.04	2	M-20		7.23	2
M-9		8.96	2	M-21		7.33	3
M-10		9.81	3	M-22		7.59	3
M-11		9.34	2	M-23		9.73	2
M-12		9.49	2	M-24		9.09	2

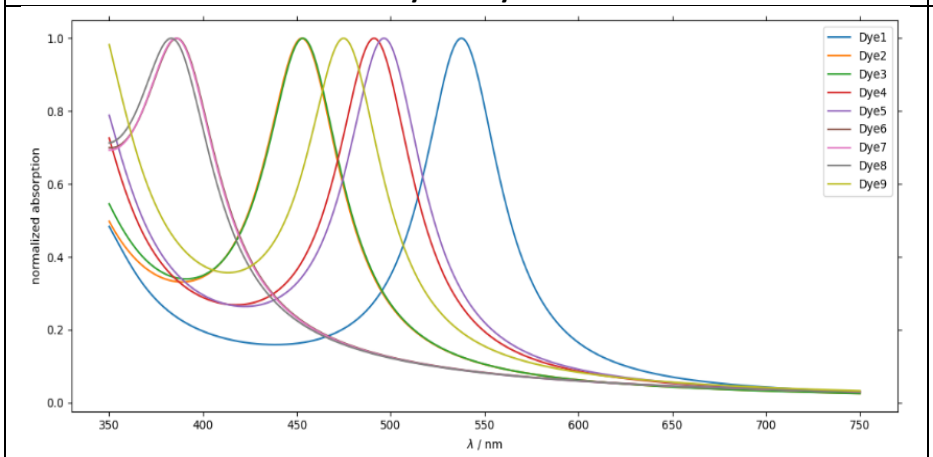
M-25		9.26	3	M-38		10.10	3
M-26		9.32	3	M-39		10.42	2
M-27		9.24	2	M-40		9.79	2
M-28		9.33	2	M-41		9.57	3
M-29		8.63	3	M-42		9.26	3
M-30		9.96	3	M-43		9.91	2
M-31		9.69	2	M-44		9.72	2
M-32		9.92	2	M-45		9.54	3
M-33		9.11	3	M-46		9.54	3
M-34		9.42	3	M-47		10.57	2
M-35		10.69	2	M-48		9.95	2
M-36		10.63	2	M-49		9.29	3
M-37		9.66	3	M-50		9.72	3

M-51		9.78	2	M-64		7.82	1
M-52		9.54	2	M-65		8.28	4
M-53		9.12	3	M-66		8.88	4
M-54		9.75	3	M-67		9.81	1
M-55		8.95	1	M-68		9.03	1
M-56		8.70	1	M-69		8.23	4
M-57		8.84	1	M-70		9.47	4
M-58		7.77	1	M-71		8.64	1
M-59		10.02	1	M-72		7.55	1
M-60		9.26	1	M-73		8.1	4
M-61		9.04	4	M-74		9.08	4
M-62		10.07	4	M-75		10.64	3

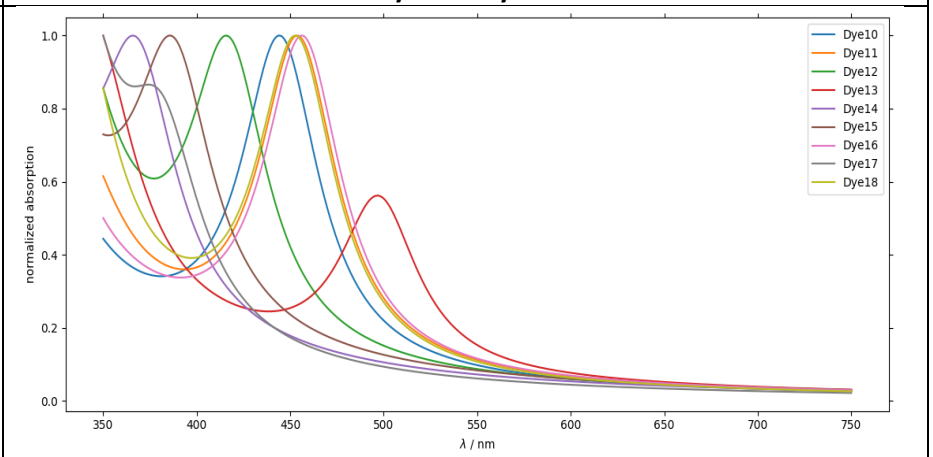
M-63		9.29	1	M-76		11.31	2
M-77		10.53	2	M-79		10.59	2
M-78		9.97	3	M-80		11.04	2

8.3 Espectros de Absorção Teórico

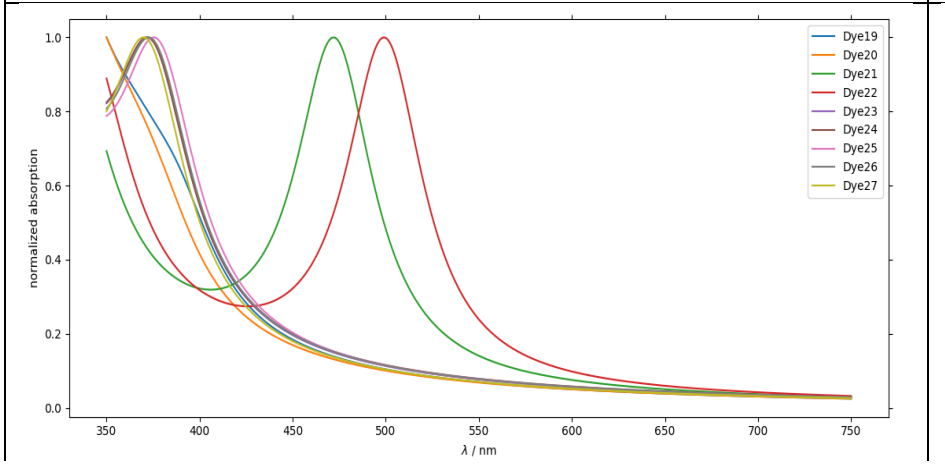
Dye1 – Dye9



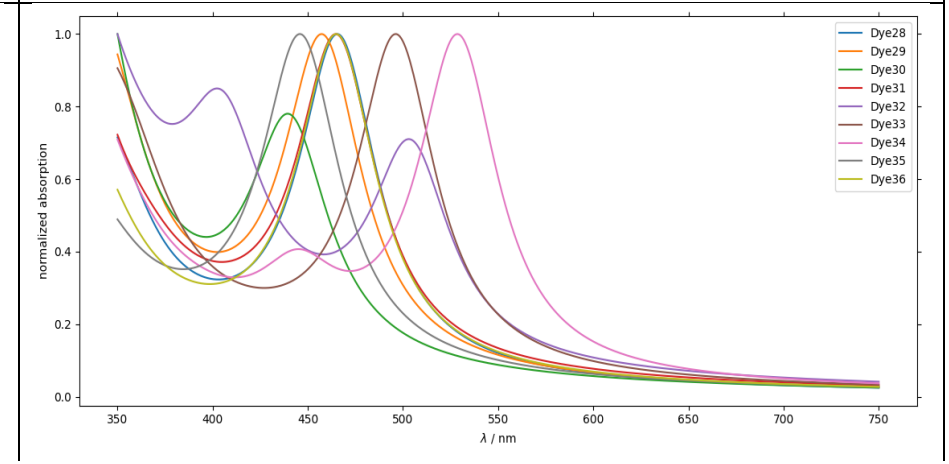
Dye10 – Dye18



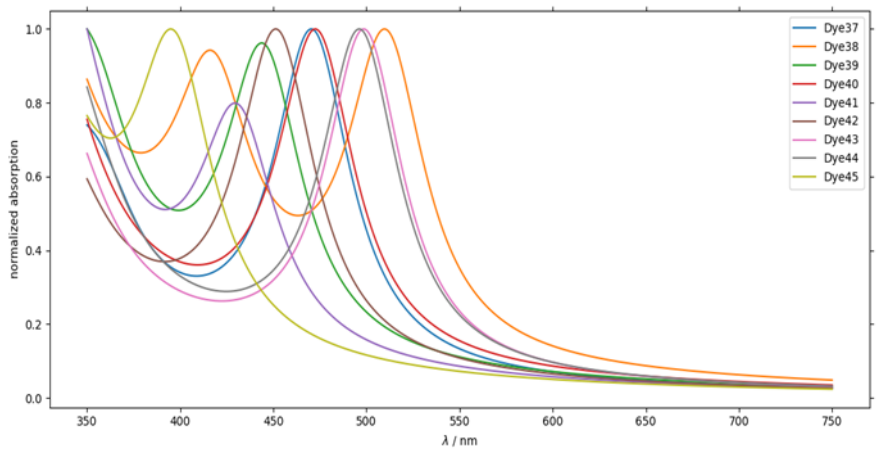
Dye 19 – Dye27



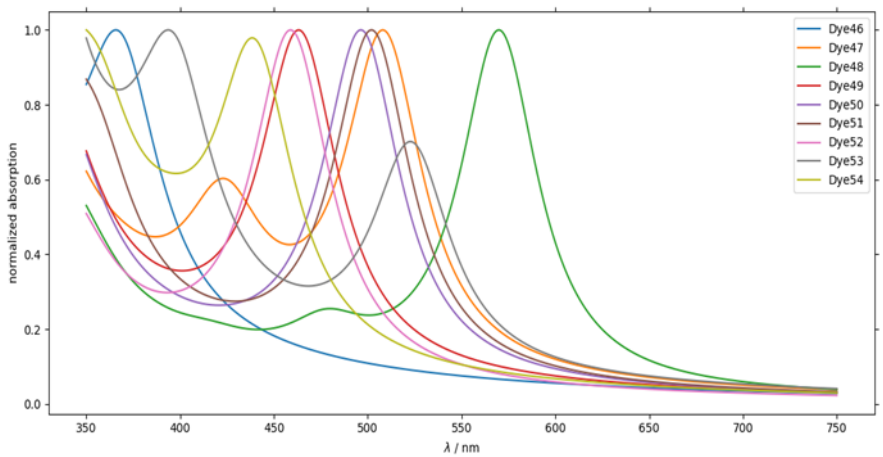
Dye28 – Dye36



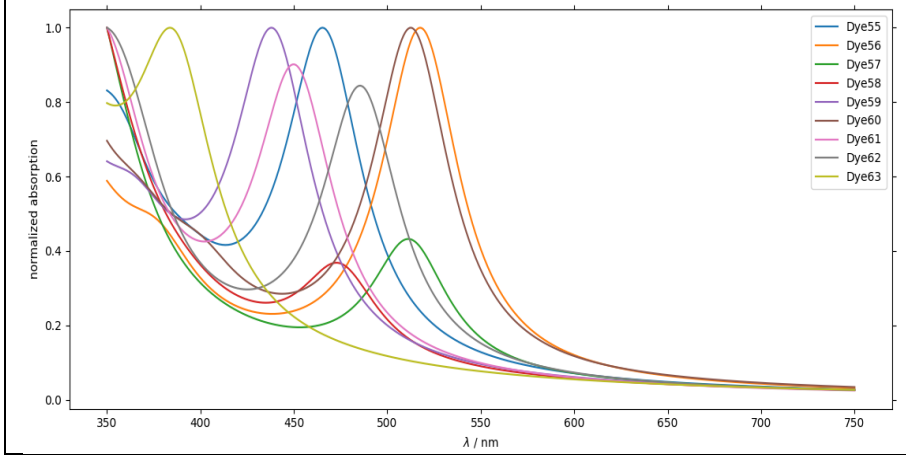
Dye37 – Dye45



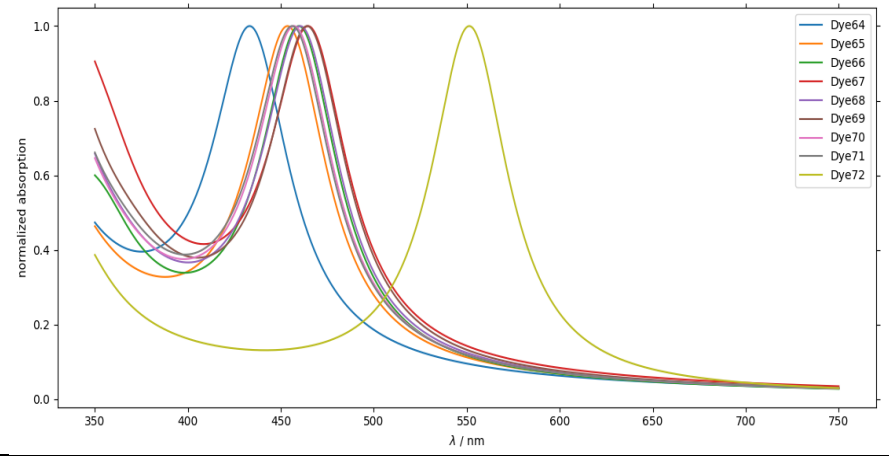
Dye46 – Dye54



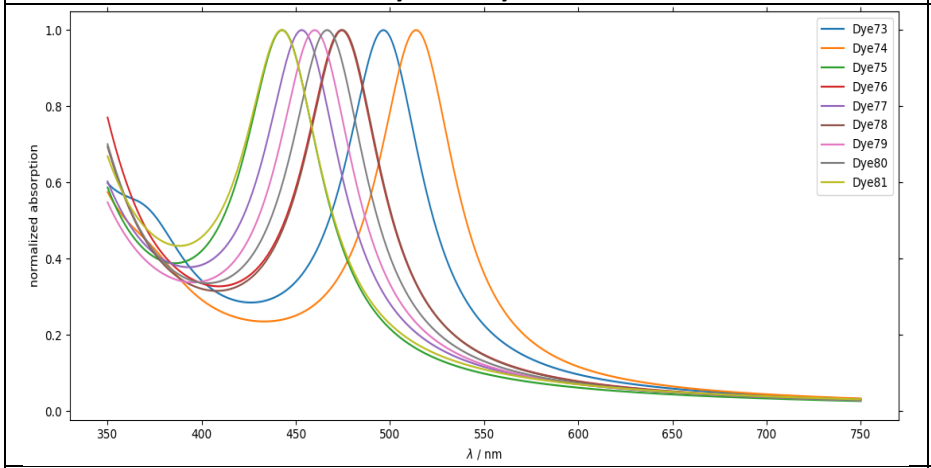
Dye55 – Dye63



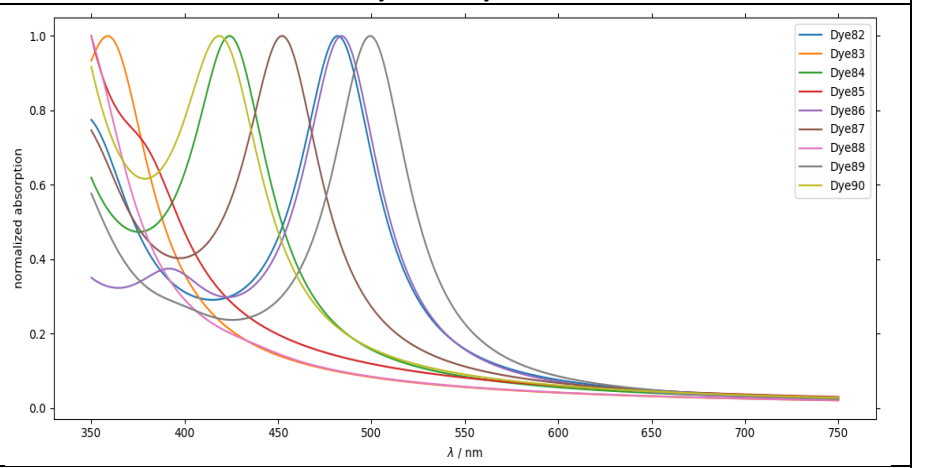
Dye64 – Dye72



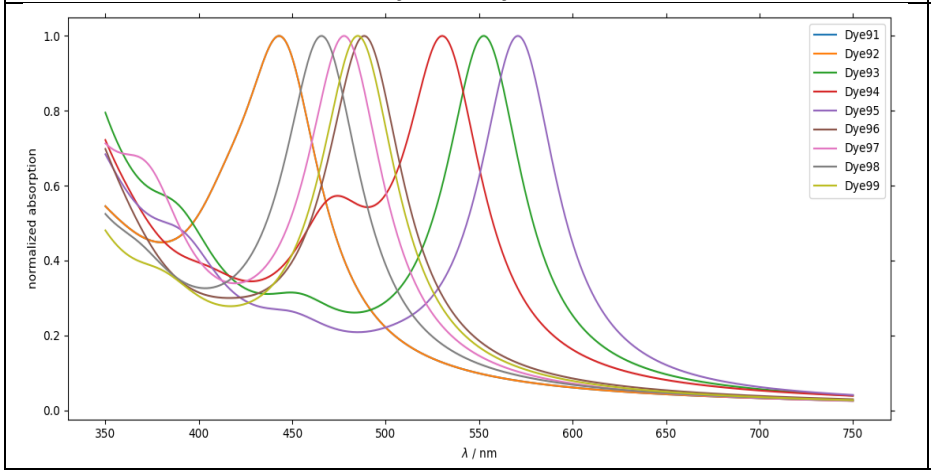
Dye73 – Dye81



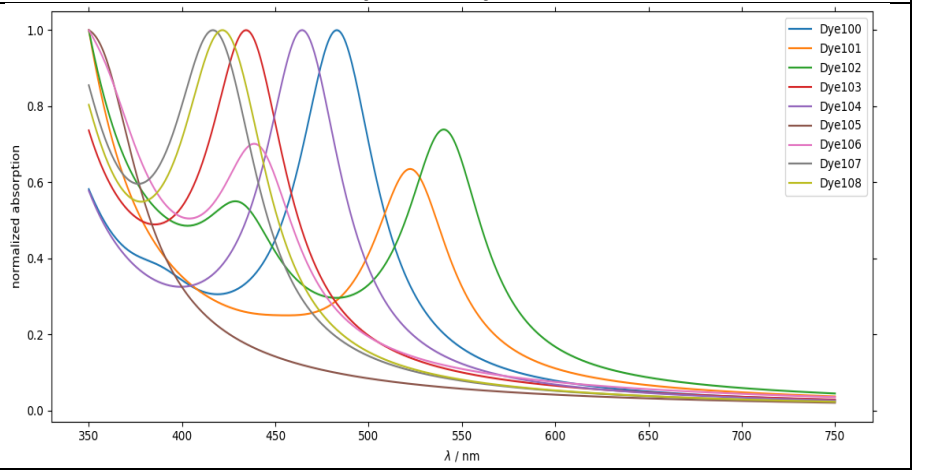
Dye82 – Dye90

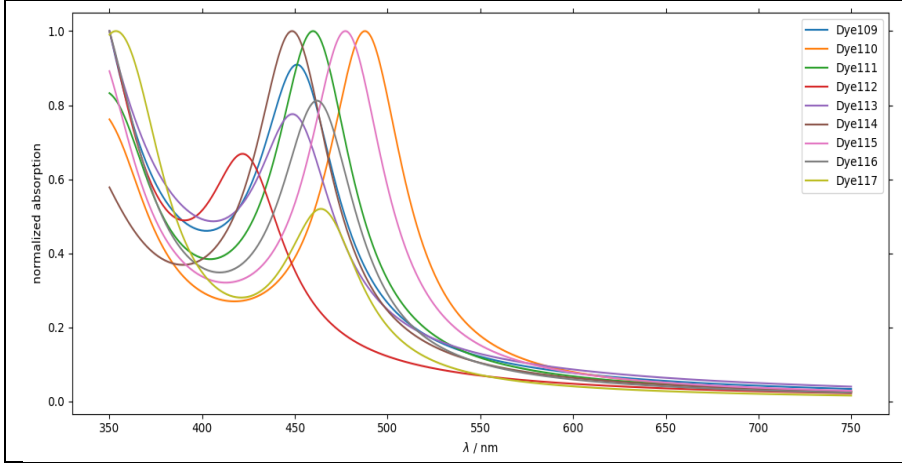
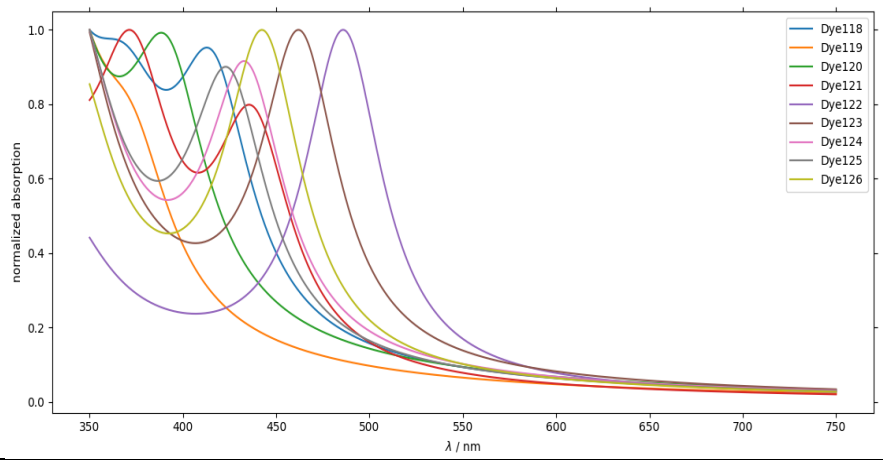


Dye91 – Dye99

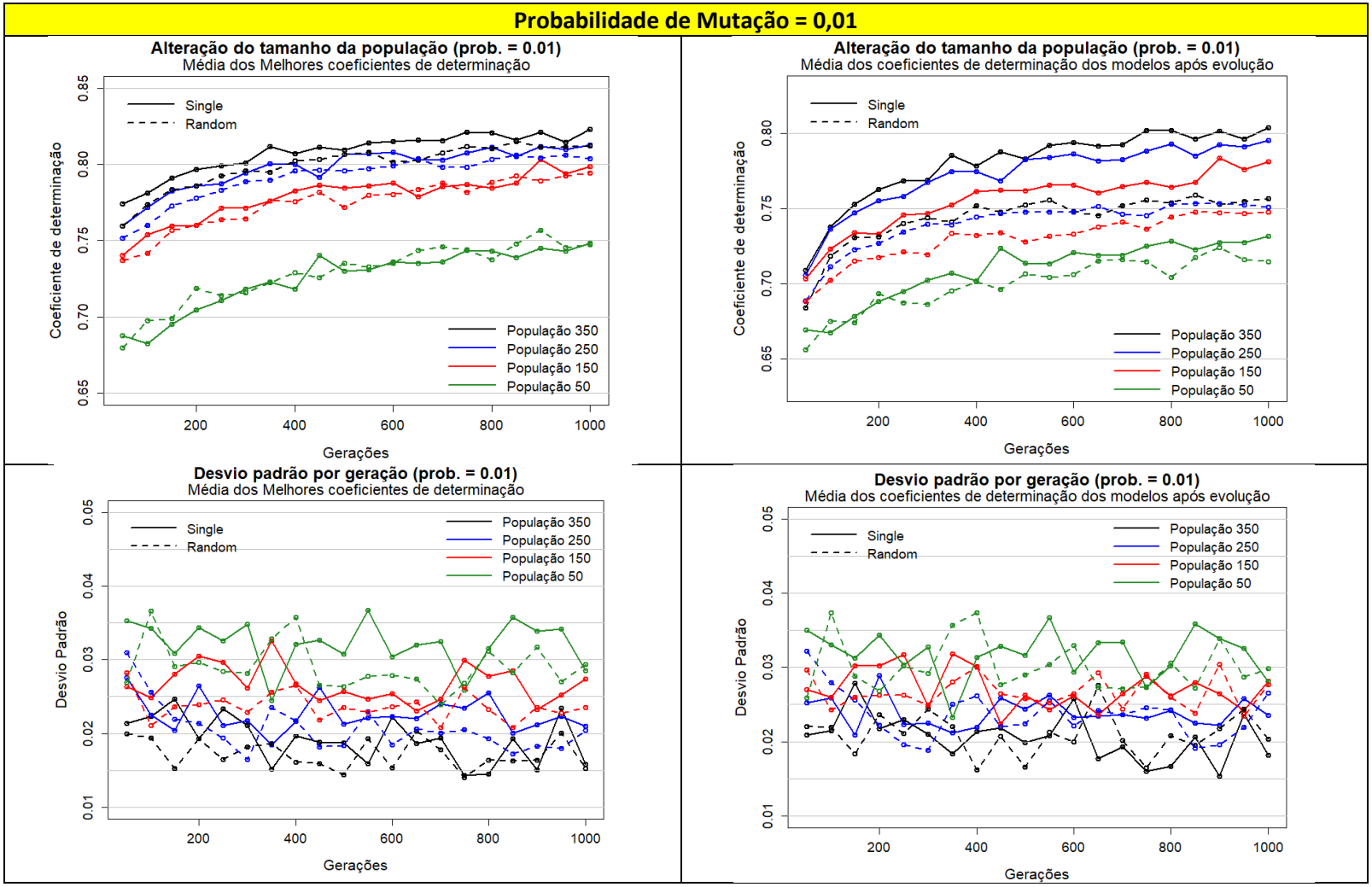


Dye100 – Dye108

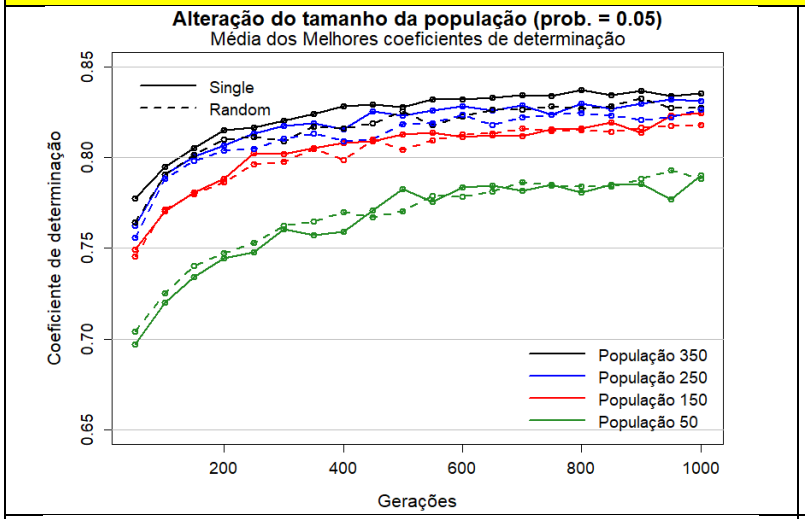


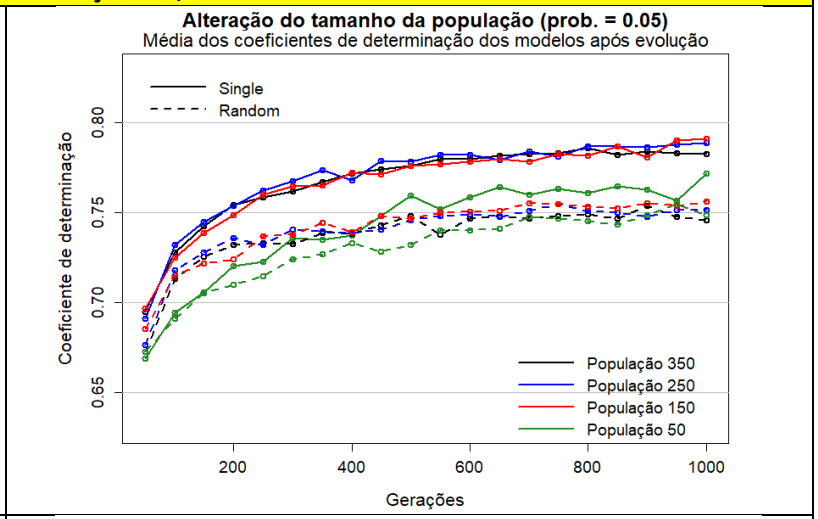
Dye109 – Dye117	Dye118 – Dye126
	

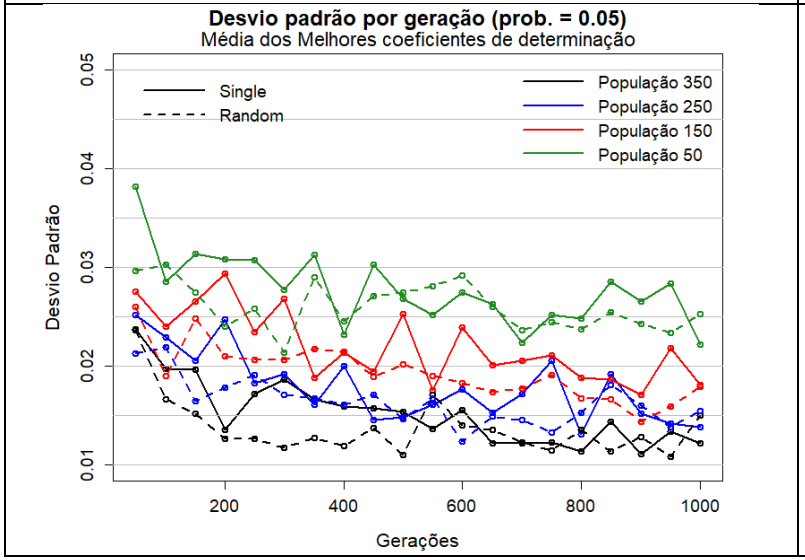
8.4 Gráficos para parametrizar o GA

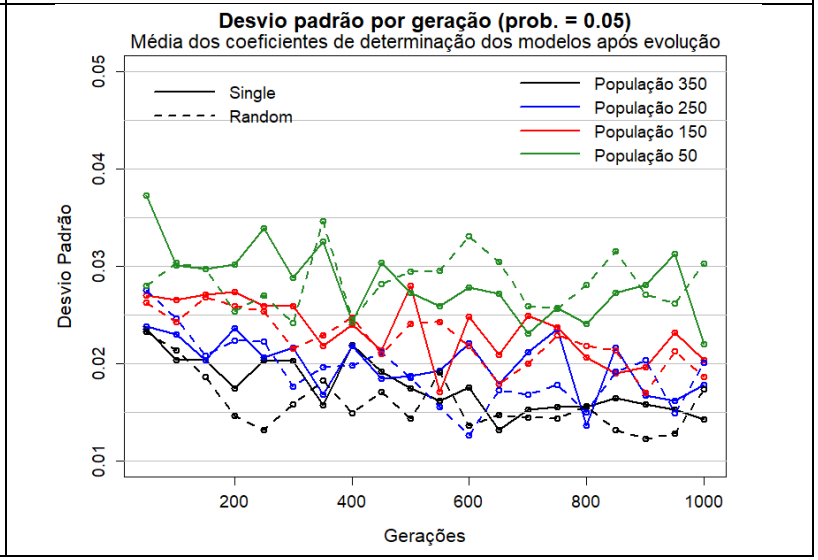


Probabilidade de Mutação = 0,05

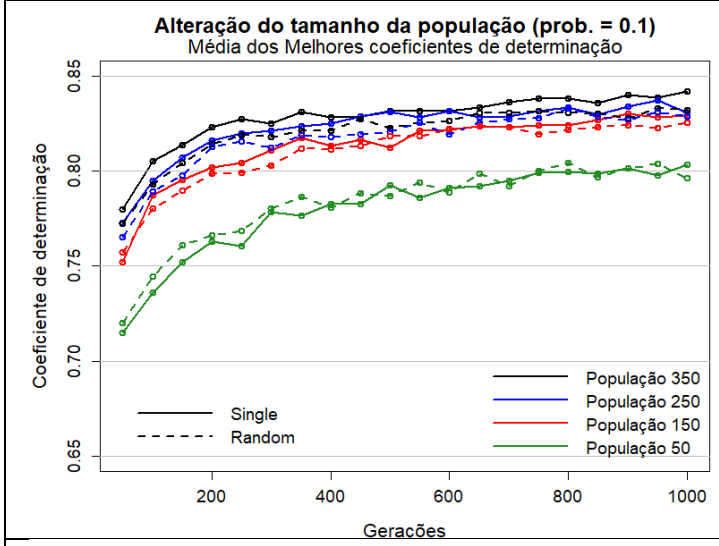


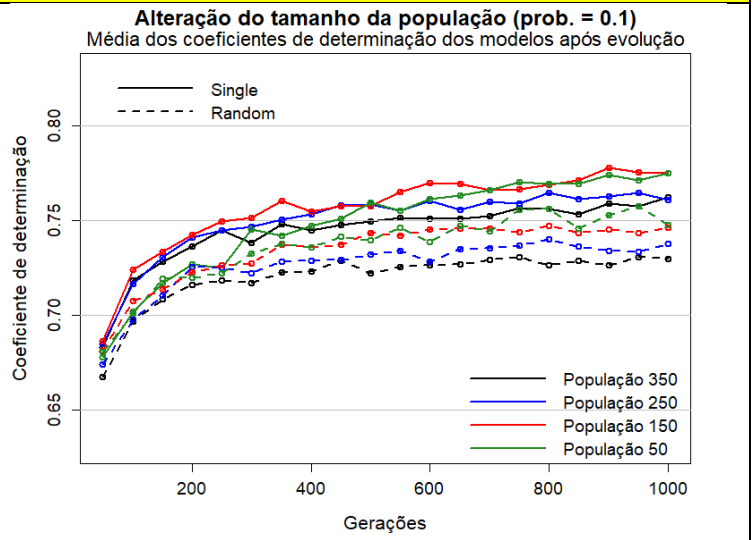


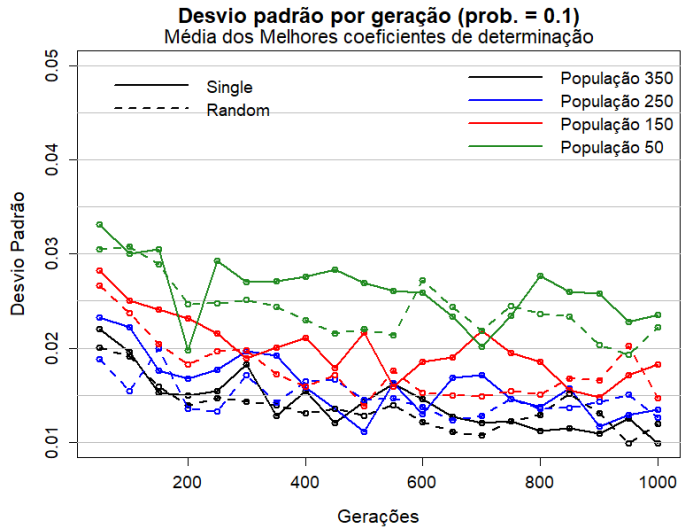


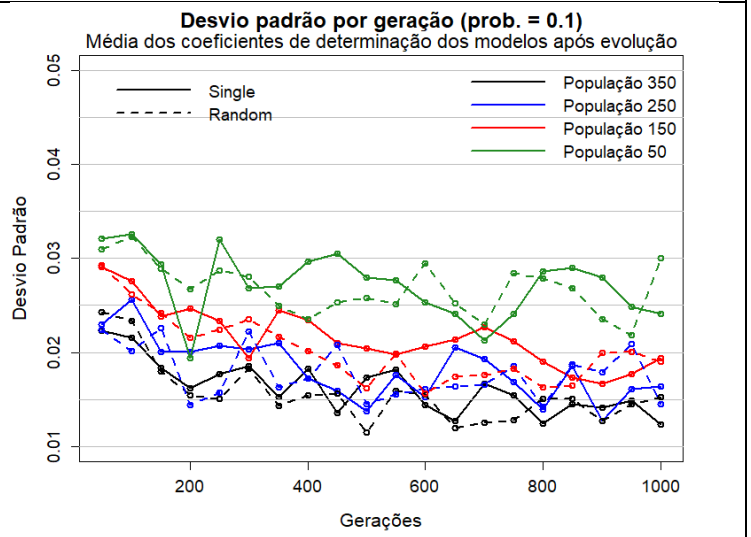


Probabilidade de Mutação = 0,1

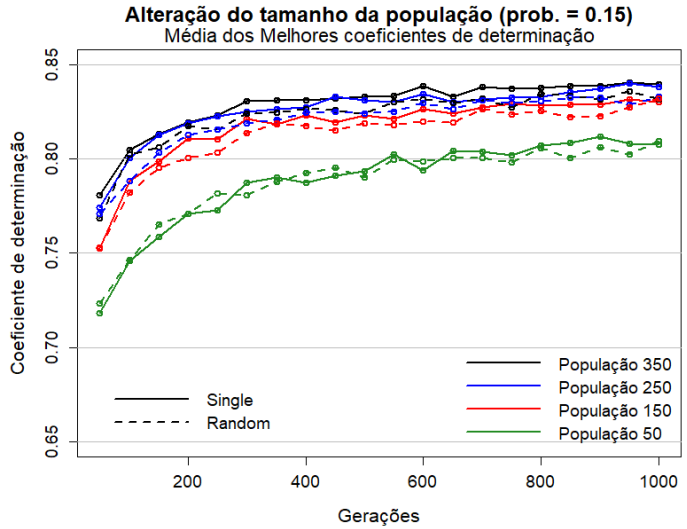


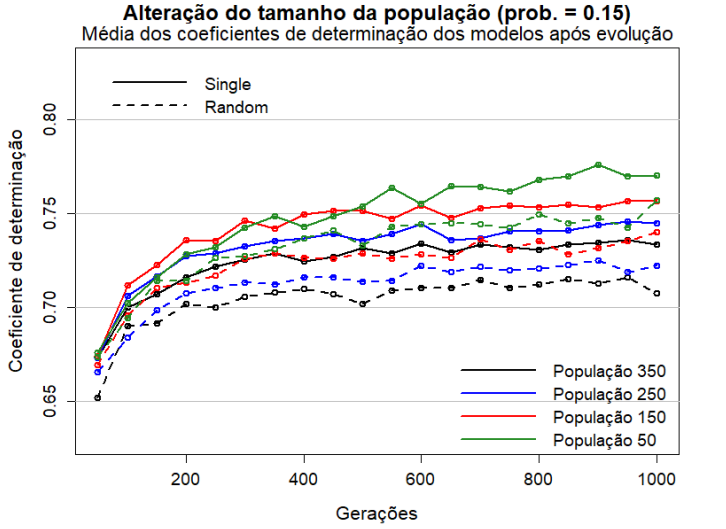


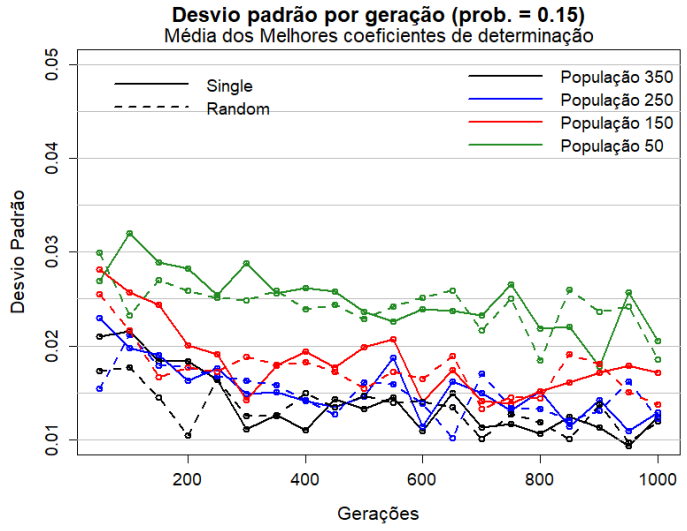


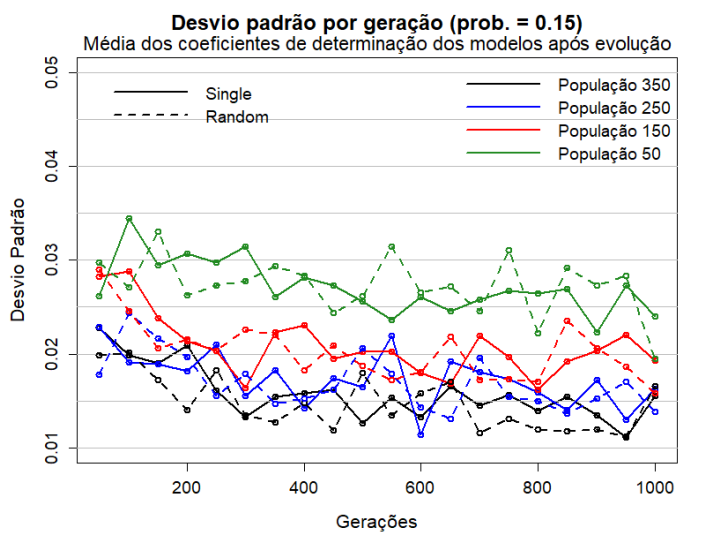


Probabilidade de Mutação = 0,15

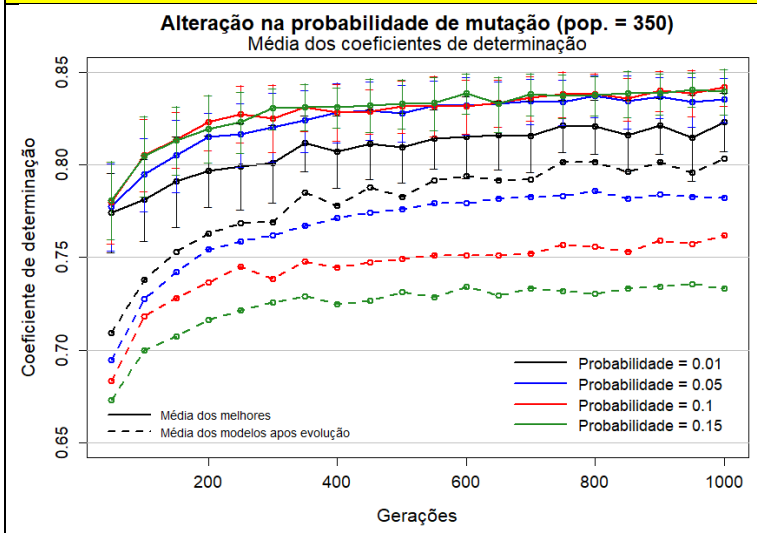


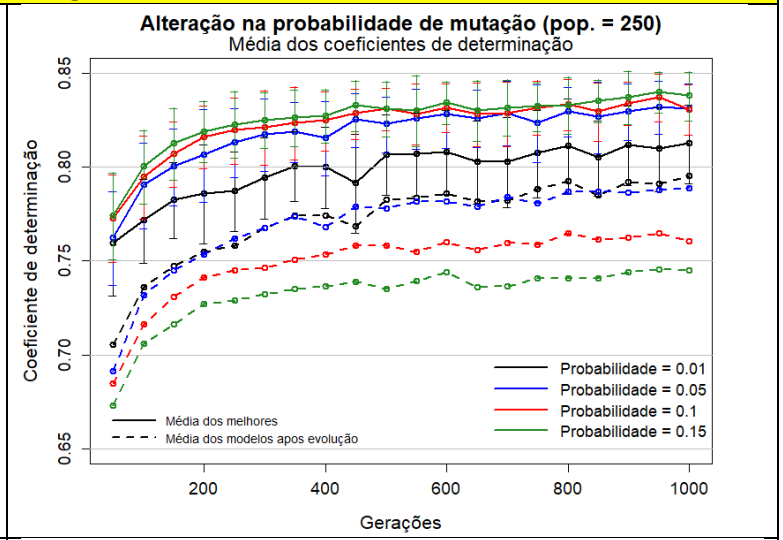


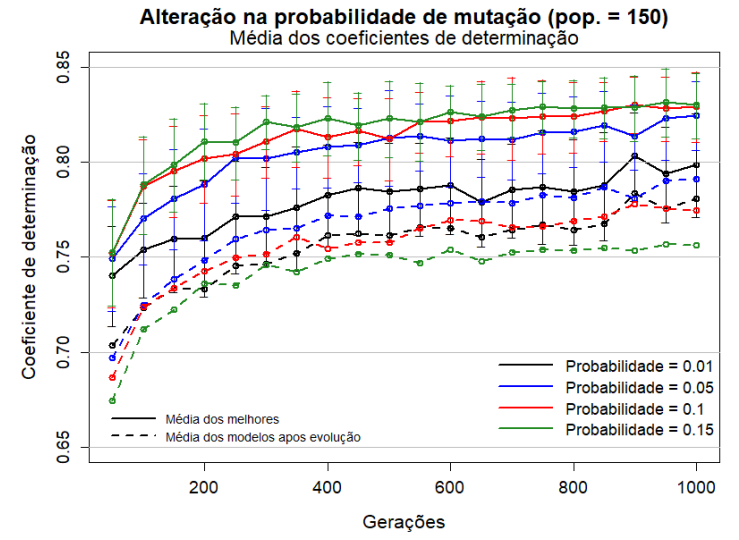


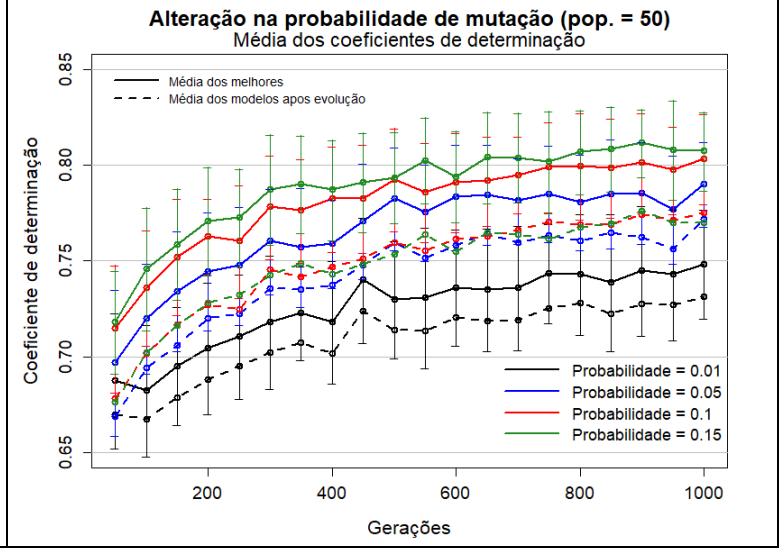


Crossover = single



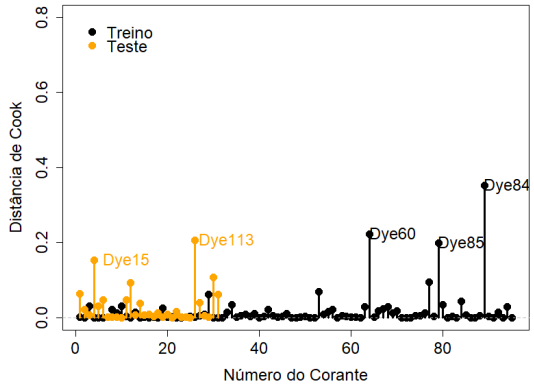




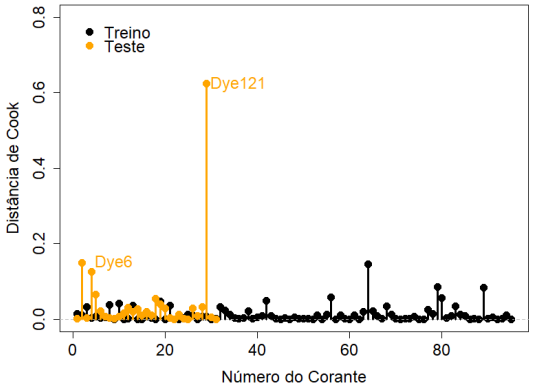


8.5 Distância de Cook

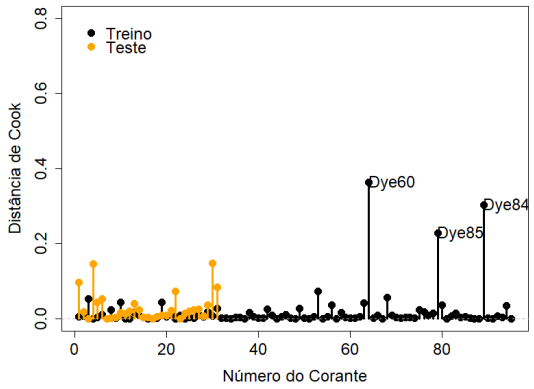
Modelo 1



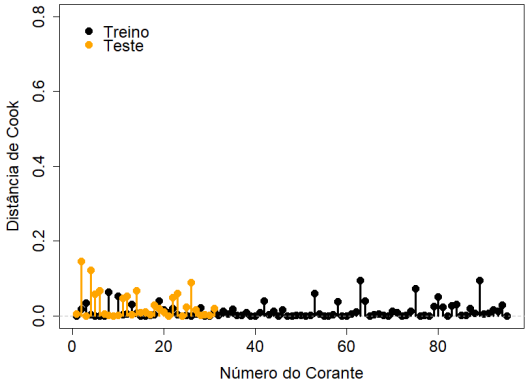
Modelo 2



Modelo 3



Modelo 4



8.6 Script para obtenção dos modelos via GA

```
# Bibliotecas
require(Metrics)
require(dplyr)
require(gaselect)
library(stringr)


## Criando os dataframes de saída

# Fórmulas
colec_formulas <- data.frame(matrix(ncol = 1, nrow = 0))
colnames(colec_formulas) <- "Formula"

# Métricas
colec_metricas <- data.frame(matrix(ncol = 11, nrow = 0))
colnames(colec_metricas) <- c("R2", "R2test", "AdjR2", "RMSE", "RMSEpred",
                              "MAE", "MAEpred", "Est_F", "DF1", "DF2", "SEED")

# Prováveis Modelos
colec_prov_model <- data.frame(matrix(ncol = 12, nrow = 0))
colnames(colec_prov_model) <- colnames(colec_metricas)


## Definindo nº de ciclos (10000 para o GA e 2000 para o bestsubsets)

nciclo <- xxx ## Definir o núemro de cíclos

for (ciclo in 1:nciclo) {

    randseed <- as.integer(runif(1, min = 0, max = 100000))
    VarGL <- as.matrix(Treino[, 5:length(Treino)])
    ctrl <- genAlgControl(populationSize = 250, numGenerations = 500, minVariables = 5,
                          maxVariables = 12, mutationProbability = 0.05)
    LimEvalu <- evaluatorLM(statistic = "r.squared")

    ## Realizando o GA para o PCE (Treino$PCE)
    result <- genAlg(Treino$PCE, VarGL, control = ctrl, evaluator = LimEvalu, seed = randseed)

    # Obtendo matrizes com os modelos da iteração
    subsetsGA <- as.data.frame(result@subsets)
    row.names(subsetsGA) <- colnames(VarGL)

    #  Matriz de métricas para os modelos da iteração
    Matriz_Formulas <- data.frame(matrix(ncol = 1, nrow = 0))
    colnames(Matriz_Formulas) <- "Formula"

    # Matriz de VIF não aceitaveis para um modelo
    ValorVIFNA <- data.frame(matrix(ncol = 1, nrow = 0))

    # Matriz de p-valores nao aceitaveis para um modelo
    ValorPNA <- data.frame(matrix(ncol = 1, nrow = 0))


    # Matriz com as metricas dos modelos
    Matriz_Metricas <- data.frame(matrix(ncol = 11, nrow = 0))
    colnames(Matriz_Metricas) <-  c("R2", "R2test", "AdjR2", "RMSE", "RMSEpred",
                                    "MAE", "MAEpred", "Est_F", "DF1", "DF2", "SEED")
```

```
# Matriz com as formulas dos modelos
Matriz_Formulas <- data.frame(matrix(ncol = 1, nrow = 0))
colnames(Matriz_Formulas) <- "Formula"

 #Gera um vetor de comprimento 15, mas com valores nulos. Vetor usado para identificar as
 variáveis no modelo
Var_mod <-  rep("", 15)

### Laço que obtém as metricas para todos os modelos gerados
for (k in 1:length(result@rawFitness)){
     i <- 1
     str_conc <- "1"  #String usada na obtenção da expressão

     # Esse laço devolve a expressão do k-ésimo modelo dentro da iteração
     for(j in 1:nrow(subsetsGA)) {

       if(subsetsGA[j, k] == "TRUE") {
         Var_mod[i] <- row.names(subsetsGA[j, ])
         str_conc <- str_c(str_conc, Var_mod[i], sep = " + ")
         i <- i + 1
       }

     }
     formula <- as.formula(str_c("PCE ~ ", str_conc)) # Formula do modelo k
     Matriz_Formulas[k, ] <- str_c("PCE ~ ", str_conc)

     #  Treinando e obtendo as métricas (R2; Q2; R2test; RMSEs; MAEs)
     #   Modelo linear
     lmmod <- lm(formula, data=Treino)
     sum_lmmod <- summary(lmmod)

     # Métricas do conjunto de Treino
     ## - R2, AdjR2, F, RMSE, MAE, p-valor e VIF
     R2 <- sum_lmmod$r.squared  # R^2 do modelo
     AdjR2 <- sum_lmmod$adj.r.squared  # R^2 ajustado
     Est_F <- sum_lmmod$fstatistic  # Estatística F
     RMSE <- rmse(lmmod$fitted.values, Treino$PCE) # RMSE
     MAE <- mae(lmmod$fitted.values, Treino$PCE)  # MAE
     VetorP <- sum_lmmod$coefficients[, 4]  # valor p das variáveis
     VetorVIF <- vif(lmmod)  #VIF das variáveis


     # Métricas do conjunto de Teste
     ## - R2Pred, RMSEpred e MAEpred

     numtest <- sum((Teste$PCE - predict(lmmod, newdata = Teste))^2)
     dentest <- sum((Teste$PCE - mean(Treino$PCE))^2)
     R2test <- 1 - numtest/dentest
     RMSEpred <- rmse(predict(lmmod, newdata = Teste), Teste$PCE)
     MAEpred <- mae(predict(lmmod, newdata = Teste), Teste$PCE)

     ## - Filtrando os modelos
     # Identificando VIF >= 5
     ValorVIFNA <- as.data.frame(which(VetorVIF >= 5))

     # Identificando p-value > 0.05
     ValorPNA <- as.data.frame(which(vetorP > 0.05))

     if(nrow(ValorPNA) == 0){
       if(nrow(ValorVIFNA) == 0){
```

```
            if(length(which(colec_formulas == Matriz_Formulas[k, ])) == 0){
              Matriz_Metricas[k, ] <- c(R2, R2test, AdjR2, RMSE, RMSEpred, MAE, MAEpred,
                                        Est_F, randseed)
              colec_formulas <- rbind(colec_formulas, Matriz_Formulas[k, ])
              colec_metricas <- rbind(colec_metricas, Matriz_Metricas[k, ])
              Completo <- cbind(colec_metricas, colec_formulas)
              if(R2 >= 0.6){
                if(R2test >= 0.5) {
                  colec_prov_model[(nrow(colec_prov_model) + 1), ] <- cbind(Matriz_Metricas[k,],
                                                                            Matriz_Formulas[k, ])
                  write.csv(colec_prov_model, "Colecao_Provaveis_modelos.csv")
                }
              }
            }
          }
        }

        # - Printando progresso

        n <- ncol(subsetsGA)
        progresso1 <- (k/n)*100
        progresso2 <- (ciclo/nciclo)*100
        print(str_c(k, "/", n, " (", progresso1, "%) - Ciclo: ", ciclo, "/", nciclo,  "(", progresso2, "%) -
              nº de P.M.:", nrow(colec_prov_model)))
    }

    write.csv(colec_formulas, "Colecao_formulas.csv")
    write.csv(colec_metricas, "Colecao_metricas.csv")
    write.csv(Completo, "Conjunto_Completo.csv")
}
```