DESIGN OF ET MALDI MATRICES ASSISTED BY ARTIFICIAL INTELLIGENCE (AI) FROM THEORETICAL AND EMPIRICAL DATA

A thesis submitted to the Faculty of Science in partial fulfillment
of the requirements for the degree of Master in Science of
Chemistry

Carlos Andrés Padilla Jaramillo

Biologist

Advisor

Cristian Blanco Tirado, PhD.

Co-advisor

Aldo Fabrizzio Combariza Montañez, PhD.

Marianny Yajaira Combariza Montañez, PhD.

Universidad Industrial de Santander (UIS)

Facultad de Ciencias

Escuela de Química

Maestría en Química

Bucaramanga

2024

Acknowledgements

To my advisors, Cristian, Marianny and Aldo for their guidance, and experience throughout this research.

To the Universidad Nacional de San Agustín (UNSA, Arequipa, Perú) and the Universidad Industrial de Santander (UIS) for providing access to the INKARI and GUANE supercomputers, respectively.

To Parque Tecnológico Guatiguará and the Central Research Laboratory Facility at UIS for their infrastructural support.

To all the members of the GIFTEX and IN SILICO research groups and especially to Luis Miguel for his constant collaboration and valuable advice that helped me in my self-growth.

To my friends Bleidy, Evelyn and Laura. Your presence has been invaluable.

My deepest gratitude to Maria José for always providing me with strength and clarity.

To all my family, for the support, love, and advice.

Table of content

List of Figures

Listo of Tables

Abstract

Title: Design of ET MALDI matrices assisted by Artificial Intelligence (AI) from Theoretical and Empirical data

Authors: Carlos Andrés Padilla Jaramillo, Biologist

Keywords: Rational Design, MALDI MS, Artificial Intelligence, Quantum Mechanics.

The development of matrices for Matrix-Assisted Laser Desorption/Ionization Mass Spectrometry (MALDI MS) has traditionally relied on experimental efforts. Here, we propose a Goal-Directed artificial intelligence generative model, fueled by computational chemistry calculated data, to construct a chemical space optimized for Electron Transfer (ET) processes in MALDI analysis. We utilized a group of 30 reported ET matrices, subjected to structural enumeration and molecular properties prediction using semiempirical and ab initio calculations, to establish a comprehensive database comprising diverse structural and property data. Subsequently, employing a protocol of structural enumeration with 68 canonical SMILES of Bemis-Murcko (BM) fragments, we expanded the structural complexity of the initial library. This process generated 82753 compounds organized into 10 scaffold levels, with a p50 index from the Cyclic System Retrieval (CSR) curve of scaffolds of 50%. From the resulting enumerated library, a diverse subset of structures was selected using the Jarvis-Patrick clustering method. These structures, along with their associated properties measured from quantum mechanics and experimental data, were used to train a Machine Learning (ML) model to predict ionization energy (E_i) values. Subsequently, a Scoring Neural Network (SNN), coupled with our Goal-Directed generative model using Recurrent Neural Network (RNN) with Deep Learning (DL) architectures, was trained. The generative model was guided using a prior network within a Reinforcement/Transfer Learning environment. The final AI-generative model learned that structures with high unsaturation, H/C ratios under 1, and molecular weights between 100 u and 300 u are favorable for ET MALDI matrices, as well as those with few aromatic rings and zero aliphatic rings. Other molecular features were also favored. The resulting AI-generated library exhibits E_i values over 8.0 eV, akin to those of reported “good” ET MALDI matrices, indicating successful design with high synthesis accessibility scores. In conclusion, our generative model provided valuable insights into the molecular features ideal for ET MALDI compounds while generating a wide range of structurally diverse molecules within a similar molecular property space. The next critical step in this process is to synthesize a selection of these generated compounds for experimental validation and further characterization.

Resumen

Título: Diseño de matrices MALDI TE asistido por Inteligencia Artificial (IA) a partir de información teórico/experimental

Authors: Carlos Andrés Padilla Jaramillo, Biólogo

Keywords: Diseño Racional, MALDI MS, Inteligencia Artificial, Mecánica Cuántica.

El desarrollo de matrices para la espectrometría de masas por desorción/ionización láser asistida por matriz (MALDI MS) ha dependido tradicionalmente de esfuerzos experimentales. Aquí, proponemos un modelo generativo de inteligencia artificial orientado a objetivos, impulsado por datos calculados de química computacional, para construir un espacio químico optimizado para procesos de Transferencia Electrónica (ET) en el análisis MALDI. Utilizamos un grupo de 30 matrices de ET reportadas, sometidas a enumeración estructural y predicción de propiedades moleculares mediante cálculos semiempíricos y ab initio, para establecer una base de datos integral que comprende diversos datos estructurales y de propiedades. Posteriormente, empleando un protocolo de enumeración estructural con 68 SMILES canónicos de fragmentos de Bemis-Murcko (BM), expandimos la complejidad estructural de la biblioteca inicial. Este proceso generó 82,753 compuestos organizados en 10 niveles de andamios, con un índice p50 de la curva de Recuperación de Sistemas Cíclicos (CSR) de andamios del 50%. De la biblioteca enumerada resultante, se seleccionó un subconjunto diverso de estructuras utilizando el método de agrupamiento de Jarvis-Patrick. Estas estructuras, junto con sus propiedades asociadas medidas a partir de datos de mecánica cuántica y experimentales, se utilizaron para entrenar un modelo de Aprendizaje Automático (ML) para predecir los valores de energía de ionización (Ei). Posteriormente, se entrenó una Red Neuronal de Puntuación (SNN), acoplada a nuestro modelo generativo orientado a objetivos utilizando Redes Neuronales Recurrentes (RNN) con arquitecturas de Aprendizaje Profundo (DL). El modelo generativo fue guiado utilizando una red previa dentro de un entorno de Aprendizaje por Refuerzo/Transferencia. El modelo generativo final de IA aprendió que las estructuras con alta insaturación, relaciones H/C inferiores a 1 y pesos moleculares entre 100 u y 300 u son favorables para las matrices ET MALDI, así como aquellas con pocos anillos aromáticos y cero anillos alifáticos. También se favorecieron otras características moleculares. La biblioteca generada por IA resultante exhibe valores de Ei superiores a 8.0 eV, similares a los de las matrices ET MALDI “buenas” reportadas, lo que indica un diseño exitoso con altos puntajes de accesibilidad de síntesis. En conclusión, nuestro modelo generativo proporcionó valiosas ideas sobre las características moleculares ideales para los compuestos ET MALDI, al tiempo que generó una amplia gama de moléculas estructuralmente diversas dentro de un espacio de propiedades moleculares similar. El siguiente paso crítico en

este proceso es sintetizar una selección de estos compuestos generados para validación experimental y caracterización adicional.

Theoretical framework

MALDI Mass Spectrometry and Electron Transfer (ET) matrices

Mass spectrometry (MS) is a very useful chemical-analytical technique that primarily uses molecular ionization to identify chemical structures (1). MS has a high detection limit, is sensitive, selective, accurate and fast, making it an ideal technique for many applications (2, 3). The machinery and instrumentation used in MS assays consist mainly of an ionization source, a mass analyzer, and a detector (Fig. 1) (4). Pioneering MS assays were carried out with high concentrations of analytes and single molecules obtained from extensive purification processes (5). Today, MS assays are used to study macromolecules such as proteins, peptides, oligonucleotides, lipids, DNA, carbohydrates and other biologically relevant molecules at low concentrations, with highly successful results (6-8).

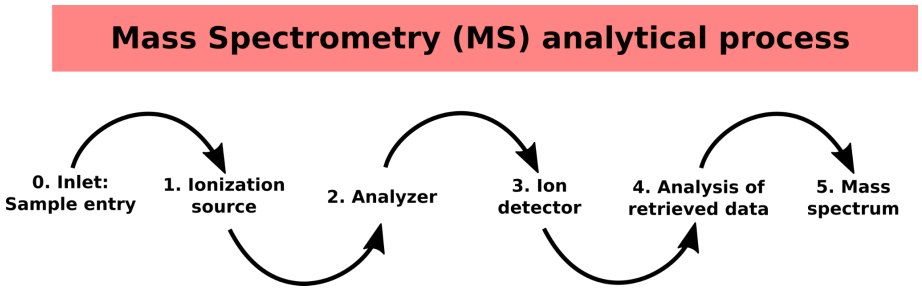


Figure 1. Mass spectrometry workflow.

The ionization method is chosen depending on the analyte structure and requires a specific research approach (2). Currently, the most popular ionization sources are electrospray ionization (ESI) and Matrix-Assisted Laser Desorption/Ionization (MALDI) (3). Other sources include electron ionization (EI), chemical ionization (CI), atmospheric pressure chemical ionization (APCI), nano-ESI, desorption electrospray ionization (DESI), laser ablation ESI ionization (LAESI), photoionization (PI), desorption/ionization on silicon (DIOS), fast atom pump (FAB), secondary ion mass spectrometry (SIMS), direct analysis in real time (DART), flow after discharge at atmospheric pressure (FAPA), dielectric barrier ionization (DBDI) and inductively coupled plasma (ICP) (9).

In particular, MALDI MS was developed in 1988 by Takana *et al.* and Karas & Hillenkamp (10,11). The MALDI laser source allows ionization of molecules of large molecular mass (>10000 u) over a wide range of concentrations (2). The analyte is co-crystallized with an organic compound called matrix (analyte-matrix mixture), which has specific physico-chemical

properties favorable for the Electron Transfer (ET) process to occur (Fig. 2). The ET process in MALDI consists of two main steps. In the first step, a laser beam (λ = 355 nm) is incident on the matrix-analyte mixture placed on a target plate. The matrix-analyte is ionized and desorbed. In the second step, the neutral analyte transfers an electron (ET process) to the ionized MALDI matrix (12). To achieve the ET process, the ionization energy (E_i) of the matrix must be higher than that of the analyte to ensure charge migration (13), among other important thermochemical aspects.

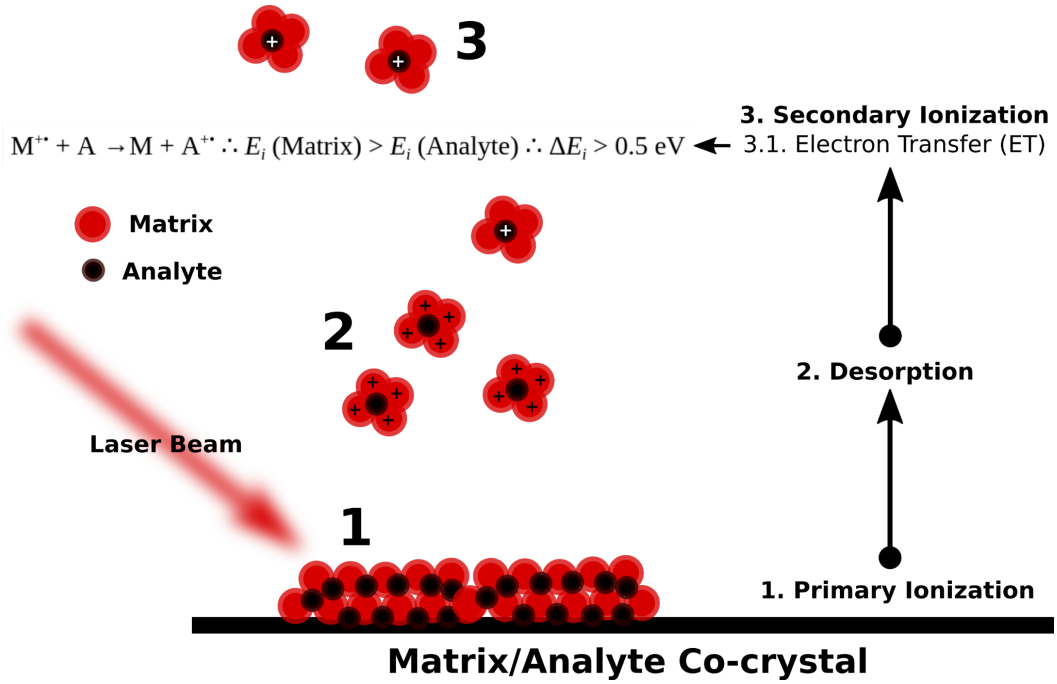


Figure 2. Scheme of Matrix-Assisted Laser Desorption/Ionization (MALDI) source.

Rational design of ET MALDI matrices

The search for new MALDI matrices with improved physico-chemical properties to analyze a wide variety of analytes is a constantly growing field of research. Most of the current research is based on experimental trial-and-error approaches, while chemical-computational perspectives are little explored (14). Rational design and synthesis of MALDI matrices must take into account physicochemical parameters and co-crystallization is an important step to improve analytical performance (15). Taken together, experimental and theoretical approaches are combined to develop matrices with improved physicochemical properties and increased efficiency in analyte detection (13, 16-21).

On the other hand, in 2008, Jaskolla and co-workers presented the first formal study of rational MALDI matrix design based on quantitative characteristics to select and validate potential MALDI matrix candidates, investigating α-cyanocinamic acid (CCA) derivatives (22). The

researchers used two experimental and one theoretical criteria to characterise CCA derivatives as matrices, with the theoretical stimulus criterion being paramount. All CCA derivatives in positive ionization mode exhibit good protonated matrix ion abundance signals, which plays a crucial role during the chemical ionization process via proton transfer to the analyte and gas-phase proton affinity. Moreover, all CCA derivatives with a free carboxylic (-COOH) functional group show a better ability for peptide incorporation into the matrix crystal lattice by forming ionic pairs with positively charged peptide functional groups. However, the CCA-Cl derivative was finally synthesized because its proton affinity value was low compared to the other proposed matrices. This proton affinity value was obtained from theoretical calculations, performing geometry optimizations with the Allinger MM2 force field and B3LYP/6-311G** levels of theory (22). Overall, these studies continue to build the theoretical and experimental basis for the design of specific MALDI arrays for the identification of a variety of analytes.

From Jaskolka *et al.* onwards, rational compound design studies continued to be carried out, increasingly including mechano-quantum calculations and using them as determining criteria for the design of MALDI matrices. Among the most recent MALDI matrix design studies found is the one published by Tammekivi *et al.* in 2021 where they studied for the first time monoaminoacridines (1-, 2-, 3-, 4-, 4- and 9-aminoacridine - 9AA) as matrices in the negative ionization mode of MALDI mass spectrometry for the analysis of small molecules in complex samples, performing a hybrid computational-experimental investigation (14). Some relevant physicochemical parameters were studied, such as crystallization morphology due to solvent effect, sample-matrix mass ratio effect, UV-Vis absorption and fluorescence profiles, lifetimes, quantum yields and PTRs, including self-protolysis. These matrices provided acceptable mass spectra of the samples and were suitable for the identification of the characteristic peaks present in the samples, especially the 3- and 4-AA matrices (14). Taken together, these studies continue to improve the theoretical and experimental knowledge for the design of specific MALDI matrices for the identification of a wide variety of analytes.

Cristancho (2016) explored the use of 2,7-dibromofluorene (FL) as a base to develop MALDI ET matrices. The combination of this base with rational substituents allowed obtaining new MALDI ET matrices with high E_i values, good molar absorptivity and physicochemical properties suitable for the gas-phase ET process. The FL-CN and FL-OCH_3 matrices were the best in the tests compared to the standard DCTB matrix (23). The rational design was based on theoretical E_i calculations of Koopmans' theorem and validation was performed by experimental tests.

Castellanos-García *et al.* (2017) evaluated phenylene vinylene (PV) systems with some chemical modifications, which are organic polymers with various photophysical applications and good physicochemical properties that allow fine control of the photophysical behaviour and good solubility for MALDI MS application (24). The investigation was a hybrid experimental-computational study. The parameters evaluated experimentally were UV-vis, quantum yields and MS assays with different targets. The parameter evaluated from a

computational methodology was the E_i, which was calculated with the Electronic Propagator Theory (EPT) using the Hartree-Fock (RHF) method and the 6-311G (d, p) basis function set. The PV matrices studied by Castellanos are good in comparison to the standard CHCA (PT matrix), DCTB (ET matrix) and DHB (ET matrix) matrices. They have lower intermolecular and interplanar distances in the solid state (less than 8 Å), improving the conjugation of HOMO and LUMO states and making the exciton formation process more favourable. In addition, the PV arrays operate at lower laser energies, which enhances laser lifetime and reduces source contamination.

Giraldo *et al.* (2018) and Pradilla *et al.* (2019a and 2019b) continued the research of Castellanos and co-workers by analyzing and improving the PV matrices proposed by Castellanos. Pradilla (2019a) explored PV derivatives with an additional modification of cyano functional group (-CN) at the vinyl bridges (α-CNPV) and variations of electron-donating groups (EDG) and electron-attracting groups (EWG) at the peripheral-p positions, concluding that α-CNPV-CH_3 and α-CNPV-OCH_3 present the best results (25). Giraldo (2018) analyzed petroporphyrins by ET MALDI-MS, using the α-CNPV-CH_3 matrix (13). The evaluation by Ramirez-Pradilla *et al*. (2019b) demonstrated that the α-CNPV-CH_3 matrix can be used for selective ionization and detection of nickel and vanadyl porphyrins present in ACN extracts and HPTLC-purified fractions of Middle East crude oils (26). Recently, Sanchez *et al.* also used α-CNPV-CH_3 and compared its use with the commercial MALDI ET matrix DCTB for identification of photosynthetic pigments from phytoplankton. Sanchez obtained good results using both matrices, however, he showed that both matrices have advantages and disadvantages for pigment analysis (27). Therefore, rational and targeted design methodologies still need to be improved.

Artificial Intelligence Generative Models to Molecular Design

The use of artificial intelligence in many areas of chemistry has been increasing since 2017, with easy access to digital frameworks or APIs for the general public, including non-programmers. The chemical areas utilizing AI include analytical chemistry, industrial chemistry, chemical engineering, physical chemistry, biochemistry, among others (28). On the other hand, the issue of investigation most relevant are molecular property prediction, reaction outcomes, nanotechnology, wastewater treatment, natural lenguaje processing, retrosynthesis, drug discovery, and molecular design.

Molecular computer-assisted design has been presented mainly in areas of pharmacology supported by Quantitative Structure Activity Relationship (QSAR) approaches, which has been a widely studied field (29). An example is with the OLED-type compounds, studies and designed using several AI methodologies, specially AI generative models (30, 31). Generative models for the development of *de novo* compounds are wide used, generally from a SMILES (Simplified molecular-input line-entry system) string molecular representation (32-35). However, the design

approach for ET MALDI matrices using AI-supported methodologies is a field that has not been explored to date.

It is important to highlight that SMILES is a notation that allows a user to represent a chemical structure in a way that can be used by the computer. This notation encodes the structure of molecules using short ASCII strings, which makes it easier for computational models to process and analyze vast libraries of chemical compounds. By converting molecular structures into SMILES strings, the researchers could efficiently input a diverse range of molecules into the generative model (36,37).

Among the most recent studies related to the development of compounds using AI approaches fed by quantum data, Kwak et al. (2022) conducted a fully theoretical study to design organic electronic compounds using a goal-directed generative model. In this study, the chemical space was expanded from some known chemical cores of interest using chemical enumeration, which allowed feeding a prior network of the generative model (38,39). Using the REINVENT algorithm, the authors created customized scoring functions to guide molecular generation in agreement with a range of property values. Additionally, the SYBA (SYnthetic Bayesian Accessibility) algorithm was used to determine the synthesis feasibility of the compounds generated by the generative model and the training library (40). One of the most important results of this paper is that the generative model learned to create structures with high MPO values in reference to the enumerated training library.

State of the Art

Development of MALDI matrices

Many research groups around the world, since the emergence of MALDI MS in 1987, have focused on improving spectral resolution and making the analysis process more effective. Most studies are focused on using increasingly specialized matrices to analyze certain types of chemical species such as low molecular weight molecules *e.g.* lipids, and/or unstable chemical species (41).

On the one hand, all research on MALDI matrix development until 2008 is considered as empirical design studies and mainly structured under qualitative criteria (*e.g.,* structural similarity between matrix and analyte). However, from 2008 with the publication of Jaskolla *et al.* the term "rational design" was introduced, which is based on several quantifiable factors. In addition, the study of families of molecules (chemical derivatives) and computational chemistry (both those based on principles of statistical thermodynamics and quantum mechanics) began to gain importance as a complement to experimental approaches and in some cases as a determining

factor in the selection of viable candidates (22). Importantly, the study of families of molecules has advantages for the development of MALDI matrices, as it is easy to observe how the substitution, inclusion and/or mixing of specific functional groups within the same structure fluctuates the physicochemical behaviour of the matrices (14,16).

One of the latest studies published on the rational design of MALDI matrices was in 2021, where Monoamino Acridines (AA-) molecules were designed and evaluated in negative ion mode ("proton sponge" matrices), addressing the analysis of physicochemical factors. such as pKa(s), UV-vis, phosphorescence, quantum yield (φ), co-crystallization, protonation pathways and autoprotolysis. This study is interesting and very novel since the computational and experimental approaches have the same relevance to evaluating these new candidates for MALDI matrices (14). As a result of this research, the authors determined that the 3- and 4-AA matrices were the most suitable for the study of rosin resin, stearic acid and dyes.

For its part, our Theoretical and Experimental Physicochemistry Research Group (GIFTEX) at the Industrial University of Santander (UIS) has been developing MALDI matrices for more than ten years, designing and synthesizing *de novo* structures based on fluorene nuclei, fluorene bromide and phenylvenylenes; research that has promoted the study of a great diversity of chemical species in a rational and directed manner (10-13). Specifically, GIFTEX has delved into developing MALDI MS protocols and Electron Transfer (ET) matrices to identify porphyrins, petroporphyrins, vanadyl porphyrins, pigments from phytoplankton species, etc. Among the ET matrices that have been implemented are some commercial ones such as DCTB and others developed "*in-house"* such as $\boldsymbol{\alpha}$-CNPVs (26, 42). The results they have obtained by implementing their own matrix systems to analyze this type of analytes have been satisfactory. Currently, our group is addressing the study of more matrix systems based on Fluorene, Carbazole and Triphenylamine nuclei, with which it is planned to develop MALDI matrices as Donor-Acceptor capability to analyze a wide spectrum of chemical species. This information has not been published yet.

Computational design powered by artificial intelligence for the development of compounds

According to our literature review, to date, there are no MALDI matrix development studies of particular interest using exclusively *in silico* or *de novo* methodologies. However, the methodologies and algorithms created to develop *de novo* compounds have advanced since the beginning of 2000 and excellent results have been obtained to design compounds that are viable in synthesis and efficient for the activity for which they were intended. It is important to highlight that all of these studies are actively implementing algorithms powered by artificial intelligence (28-34,43-45).

This new methodology for developing compounds generates large chemical structures by exploring the chemical space of possibilities within a framework of base structures. From the generated structures, the properties of said compounds begin to be theoretically evaluated, generally using quantum mechanical approaches (29). This way of designing and proposing compounds is effective, however, it presents several limitations such as that not all the structures generated will be useful for the initial design purposes, or not all will be chemically viable structures for synthesis and implementation.

Over the years, these computational design drawbacks have been overcome by including structure generation algorithms, based on scoring functions and symmetric numbering schemes, which allow ranking the viability and structural logic of the new ones generated. Genetic algorithms are generally used for these purposes (30,32-34,45). On the other hand, another factor considered is the multiparametric optimization (MPO) functions that allow the structure and property correlation algorithm to be told which of all the properties and structures are favorable for the design purposes, through a normalization of 0. to 1 of the physical and chemical parameters of interest. In this way, the algorithm is told how favorable or unfavorable a chemical property or their correlation is (global MPO function).

All of these processes, since they handle a large amount of data and are iterative, must be enhanced by artificial intelligence, implementing generative models. Generative models based on a library of initial structures and a range of values of initial physicochemical parameters of interest develop novel and improved compounds. That is, from a group of base data (training group for the algorithm), compounds are generated that are faithful to the chemical-structural and multiparametric characteristics of the training data group, thereby guaranteeing the creation of structures. that will meet the design and efficiency purposes for the intended activity.

Currently, it is still in the process of coupling the parameter of chemical stability and synthesis efficiency to the generative algorithms, which are generally factors evaluated in a subsequent analysis process using the “Bond Dissociation Energy” (BDE) as a stability index and as measure of structural similarity the Tanimoto distance metric (38). Regarding MALDI Electronic Transfer matrices, which is our interest, the parameters that will be taken into account are ionization energies, electronic affinities, UV-vis absorbances and quantum yield, according to the state of development regarding MALDI ET MS compounds.

Introduction

Matrix-assisted laser desorption/ionization mass spectrometry (MALDI MS) is a versatile technique for analyzing molecules with various functionalities over a broad mass range. (17,21,25,27,46-50) Analytically, MALDI-MS is appealing for its capacity to handle complex samples with minimal susceptibility to contaminants, offer rapid analysis times, and employ straightforward sample preparation protocols. (20,51,52) Also, the technique's versatility stems from the numerous potential ionization pathways dictated by the physicochemical properties of both the analyte and the matrix.

The analysis of low molecular weight and labile compounds (LMWC) presents a challenge for traditional MALDI MS due to the limited availability of suitable matrices. (27,41,53,54) Challenges hindering accurate LMWC identification and spectral interpretation include a low signal-to-noise (S/N) ratio, often associated with analyte signal suppression by the matrix, compromising the clarity and reliability of analyte detection due to insufficient signal abundance. (2,55) Background interference (noise) from the matrix in the low (m/z) window also affects the differentiation of analyte signals, complicating compound identification. (53,54) Addressing these challenges will advance the technique, matrix design, instrument optimization, and data processing methodologies to enhance the sensitivity, reproducibility, and overall reliability of MALDI MS analysis.

The electron transfer (ET) ionization pathway in positive ion mode MALDI is commonly employed for analyzing labile, unstable, and reactive molecules. As explained by the Coupled Physical and Chemical Dynamics Model (CPCD) of MALDI, the ET reaction involves charge transfer reactions from a low ionization energy (Ei) neutral analyte molecule (A) to a high Ei matrix molecular ion (M+•) to yield an analyte molecular ion (A+•), following the mechanism below: (25,27,53,42)

$$M^{+\bullet} + A \rightarrow M + A^{+\bullet} \therefore E_i(\text{Matrix}) > E_i(\text{Analyte}) \therefore \Delta E_i > 0.5 \text{ eV} \qquad \text{(Eq. 1)}$$

For the reaction to occur, there must be an energy difference of at least 0.5 eV between the reactants, as first reported by McCarley. (56) In the CPCD model, the matrix plays a direct role in the ionization process, and therefore, its physical-chemical characteristics (E_i, proton, and electron affinity) are crucial for determining a specific ionization pathway. (57,58) The availability of commercially viable ET matrices is limited, and their performance is analyte-dependent. New ET matrices are being developed through structural modifications to well-established classic ET MALDI matrices to enhance performance and diversity. This involves adding or removing key functional groups to improve physical-chemical properties such as solubility, UV-vis absorption, E_i, and electron affinity (E_a), among others. (13,16,17,19,20-22,26,54,59) These structural modifications are selected using theoretical

models, computational simulations, and experimental data to predict and optimize the desired properties of a target molecule before synthesis. For instance, some authors postulate that designing ET MALDI matrices with π-conjugated groups that absorb wavelengths of 355 nm (Nd:YAG laser) or 337 nm (N2 laser) would facilitate the ionization process (for negative or positive ion modes) and the addition of functional groups would impact solubility modulation or enhance electron density stabilization. (54) Thus, the development of ET MALDI compounds heavily relies on "trial and error" processes, often guided solely by experimentalists' expertise. (13, 16-21, 26)

Incorporating Quantum Mechanics (QM) data may enhance the structural and physicochemical property design of these compounds. (22,26) However, this traditional design process is inefficient and costly due to the challenge of calculating, measuring, and correlating extensive physical-chemical data and structural variety. Nonetheless, Novel Material Design (NMD) methodologies accelerated by QM data and artificial intelligence (AI) facilitate the selection of new compounds with desired properties, thus avoiding design bias and focusing on the material property space. (31,60-62) Specifically, *de novo* methods generate molecules with limited or no structural and property information, particularly within the realms of Deep Learning (DL). (30,32,33,38,63,64) The most robust algorithms reported for this purpose are Goal-Directed generative models created with Recurrent Neural Networks (RNN) in a Reinforcement Learning (RL) environment. (30,35,36,60) These models are employed to perform inverse design by searching for high-scoring molecules within a chemical structural and property space from an initially chosen list of candidates. (35,38,65) Mathematically, the generative models exploit the joint probability of a molecular structure within a target bundle of properties. Typically, they start with canonical SMILES, with the option of randomization to enhance model performance, enabling exploration of vast chemical spaces. (36)

The literature describes several end-to-end workflow algorithms for molecular design, with REINVENT-AI standing out as a robust Python-based code framework for designing small molecules. (65) The latest REINVENT version, 4.1.8, facilitates molecular generation using a combination of Curriculum/Reinforcement Learning methods, including atom-per-atom design, R-group replacement, scaffold hopping, and other simple setup options, and SMILES strings. (65) However, other AI-design approaches, such as structural enumeration with fragment-based design (66-69) and scaffold-based design (69), can also support molecule generation. We used these methods for molecular generation with a generative model. Additionally, other algorithms, such as Automated Quantum Mechanical Environment (AQME) (70), were integrated into the methodology to determine molecular descriptors for predicting the E_i of our compounds, utilizing supervised predictive Machine Learning (ML) models.

Considering our experience in both utilizing and developing ET MALDI matrices, along with findings from existing literature, we emphasize the importance of having ET MALDI matrices with well-tuned E_i values for analyzing chemical species such as biological pigments from phytoplankton, porphyrins, polymers, among others, which are relevant for biotechnological and

industrial applications. For instance, analytes with E_i values under 7 eV, such as chlorophylls and carotenoids (6.8 eV < E_i > 6.2 eV), can easily form a stable radical cation upon electron transfer to a primary ion of a matrix with E_i > 7.3 eV, according to Eq. 1. (24-27,42,49,71) For analyte-matrix electron transfer in positive-ion mode MALDI, compounds such as trans-2-[3-(4-tert-butylphenyl)-2-methyl-2-propenylidene]malononitrile (DCTB) and α-Cyano-Phenylenevinylenes (α-CNPV-CH3), with 8.5 eV < E_i > 7.5 eV, have been employed. (26,27,42,49) In this context, our deep generative model, combined with quantum mechanics and experimental data of topological and molecular features, was utilized to search for structures with E_i values falling within the range of 7.5 eV and 8.5 eV, based on our previous experience with reported/designed ET MALDI compounds. Key topological considerations, such as aromatic rings and molecular weight, were incorporated into the design process, (54) along with electron-withdrawing and -donating groups. Additionally, properties such as the Octanol/Water partition coefficient (SLogP) and Topological Polar Surface Area (TPSA), as well as a synthetic accessibility score, (72) were considered relevant for the design process. These insights guided the generation of compounds with potential applications in ET MALDI MS assays, particularly for analyzing labile and unstable LMWCs. Our results demonstrate the successful generation of compounds with E_i values within the desired range, characterized by specific molecular features such as molecular weight between 100 u and 350 u, 2 to 3 aromatic rings, zero aliphatic rings, low SlogP values, and medium TPSA values.

Research Question

Can we develop a methodology based on generative models, powered by neural networks and fed with theoretical/experimental information, for the rational design of ET MALDI matrices for the specific analysis of labile and thermally sensitive chemical species?

Hypothesis

Design effective Electron Transfer (ET) MALDI matrices to analyze labile and thermally stable species, through the correlation of optoelectronic properties obtained with mechano-quantum calculations.

Objectives

General

- To develop a goal-directed generative AI model to propose specific chemical structures for Electronic Transfer (ET) processes.

Specifics

- To generate a training and evaluation library within the chemical space of interest from a common set of nuclei and substituents for MALDI ET matrices.
- To train the goal-directed generative model according to the multiparametric optimization (MPO) functions fed by theoretical/experimental data.

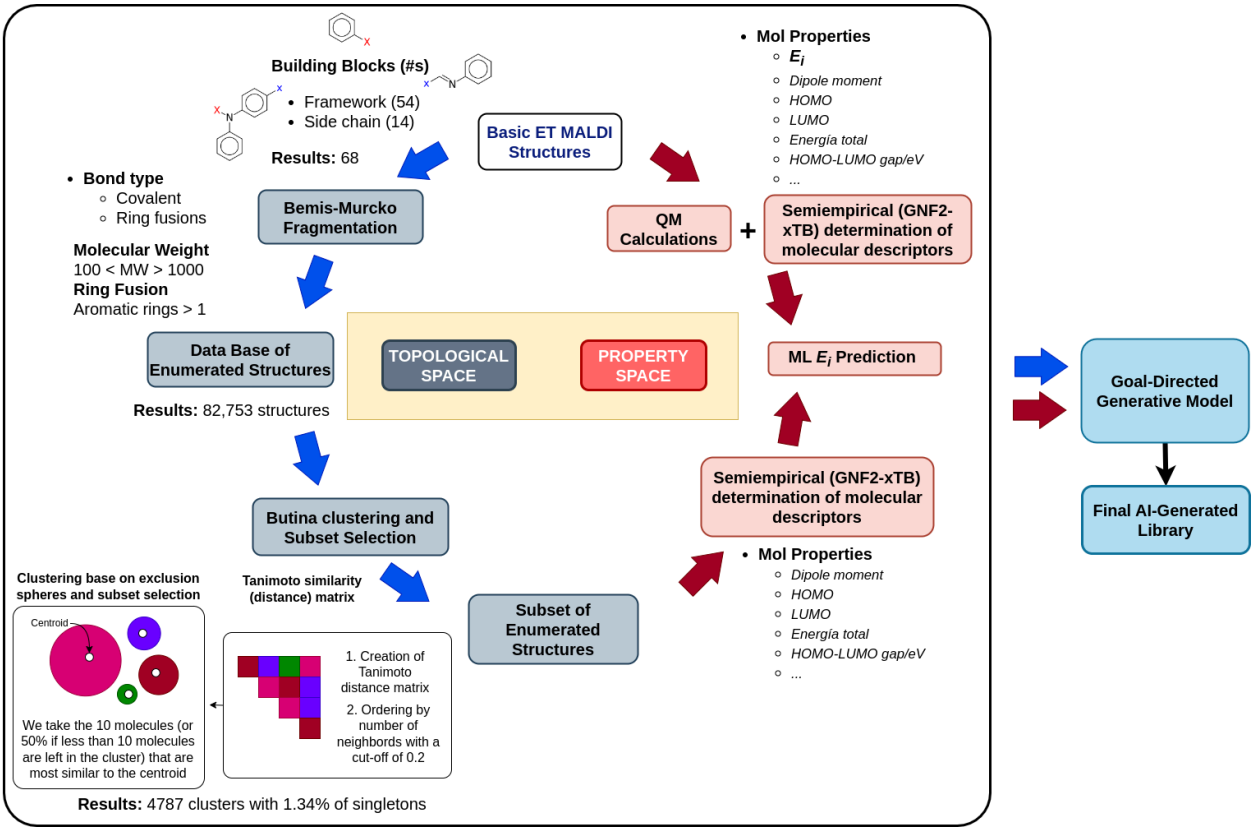
Methodology

High-Performance Computing (HPC) was utilized to perform computationally demanding calculations, including quantum mechanics jobs. These calculations were necessary for determining the ionization energies of all initial and generated compounds, as well as for deriving quantum descriptors to support the scoring network and the generative model. The HPC infrastructure afforded the requisite computational resources to efficiently execute these calculations, allowing simultaneous jobs to be conducted within reasonable timeframes. The resulting properties were employed to rank the generative model and guide molecule generation in accordance with our design objectives.

The HPC resources employed in this study were the INKARI and GUANE supercomputers, located at the Universidad Nacional de San Agustín (Arequipa, Peru), and Universidad Industrial de Santander (Santander, Colombia), respectively. INKARI comprises a head node SGI equipped with two twelve-core AMD Opteron processors running at a clock speed of 2.4 GHz, 64 GB of RAM memory operating at 1600 MHz, and 2 TB SATA 7200 RPM physical memory. Additionally, INKARI supports 36 server nodes (9x SGI), each equipped with two twelve-core AMD Opteron processors running at 2.4 GHz, 64 GB of RAM memory operating at 1600 MHz, and 2 TB Disks 7200 RPM memory. On the other hand, GUANE consists of 16 server nodes of ProLiant SL390s G7. The first 8 nodes feature 2 Intel(R) Xeon(R) CPU E5645 @2.40 GHz processors, each with twelve cores and two tasks, 104 GB RAM, 1 SAS disk of 200 GB, and 8 GPU Tesla M2075. The following 3 nodes include 2 Intel(R) Xeon(R) CPU E5645 @2.40 GHz

processors, each with twelve cores and two tasks, 104 GB RAM, 1 SAS disk of 200 GB, and 8 GPU Tesla S2050. Finally, GUANE also includes 5 server nodes with 2 Intel(R) Xeon(R) CPU E5640 @2.67 GHz processors, each with eight cores and two tasks, 104 GB RAM, 1 SAS disk of 200 GB, and 8 GPU Tesla S2050.

The following diagram shows an overview of our methodology:



Scheme 1. Methodology flowchart.

Base Topological Space and Structural Enumeration

A set of thirty-one ET MALDI matrices, including commercially available and synthesized variants (See Fig. 3), was chosen. Chemical structures were converted into unique and simple SMILES strings. (39) The internally validated chemical cores comprised triphenylamine, carbazole, and fluorene, while commercial matrices included antracene (ANT),1,5-diaminonaphthalene (1,5-DAN), 9,10-diphenylanthracene (9,10-DP-ANT), T-2-(3-(4-t-butylphenyl)-2-methyl-2- propenylidene) malononitrile (DCTB), 1,8-bis(dimethylamino)naphthalene (DMAN), tetrathiafulvalene (TTF), and terthiophene (TER). Non-commercially available compounds, α-CNPVs previously synthesized and reported as ET MALDI matrices by our group (see Fig. 3 in the molecules of reference section), were intentionally omitted from the dataset. This exclusion aimed to prevent these structures from

influencing the training or parameterization of the generative model. Instead, they were reserved for later comparison with the final library generated by our AI model. By withholding the α-CNPVs compounds during the model training process, we aim to evaluate the model's capacity to independently generate novel structures without exposure to these specific molecules. Given that the α-CNPVs matrices share similar ET scaffolds with other compounds, we anticipated that our model would produce analogous or identical structures with high scores.

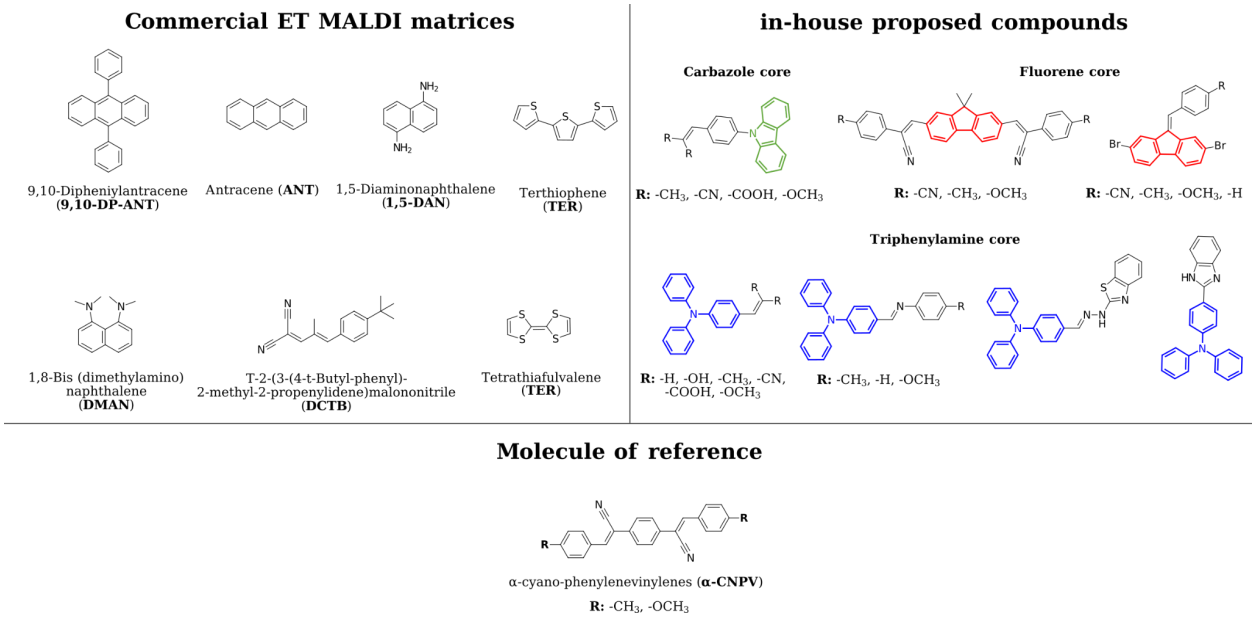


Figure 3. Commercial ET MALDI matrices and internally validated cores employed as an initial library for structural enumeration. The CNPVs served as reference molecules to validate the final AI-generated library.

The selected structures were disassembled into scaffold components consisting of aromatic systems, frameworks, and side chain building blocks, following the Bemis-Murcko method. (73) Later, the structural diversity of these building blocks was assessed using a Scaffold Network/Tree.(74) Subsequently, these building blocks were systematically combined through structural enumeration, adhering to simple chemical rules such as covalent bonds and ring fusions, with constraints set to a molecular weight of 100 u and at least one aromatic ring for the lower limit. (39,75) This process generated a broad and highly diverse valid topological space, with diversity assessed using Tanimoto indices, a Cyclic System Retrieval (CSR) curve of scaffolds, and Shannon-Entropy (SE) metrics. (76-78) The preliminary library was analyzed by grouping compounds into clusters using the Butina clustering method, employing RDKit fingerprint encodes. The threshold, python-codified into Butina algorithms, was set at 0.2. (79-81) A diverse subset of 500 compounds was screened and selected from these clusters to ensure diverse representation within the entire database. This selection process involved extracting the cluster centroid from each cluster and their ten most similar molecules.

Scoring Network and Molecular Properties

A preliminary search for conformers within the selected diverse subset and the initial ET MALDI compounds was conducted using the Automated Quantum Mechanical Environment (AQME), employing molecular mechanics potentials with the Force Fields (FF) MMFF and UFF via the RDKit module. (70) The resulting conformers were minimized using Machine Learning interatomic potentials (ANI), except for compounds containing Bromine (Br) atoms, which were optimized using the GNF2-xTB semiempirical method. (70,82) Subsequently, a set of molecular descriptors was calculated using the xTB program with the same GNF2-xTB method. Meanwhile, ORCA input files for geometry optimization were generated for each compound in our initial database to calculate the E_i using the Koopmans theorem. (83-85) The Koopmans theorem is particularly relevant for analyses within the restricted closed-shell Hartree-Fock (RHF) theory; thus, the RHF method with the basis set limit of Def2-QZVP was employed to achieve stable and minimal molecular geometry. Subsequently, the calculated and literature-retrieved E_i for the compounds were correlated with the obtained molecular descriptors. To achieve this, a Random Forest (RF) model was utilized to predict the E_i of other compounds in our diverse subset and the final AI-generated database. RF was chosen from a set of common ML models, including Neural Networks (NN), Gradient Boosting (GB), and Multivariate Linear Models (MVL), using the ROBERT-Automated ML protocols algorithm. The ROBERT input comprised 20 electronic and topological descriptors listed in Table 1, with Ei as the predicted value.

Table 1. Selected descriptors to predict Ei values using the RF model

Descriptors			
Electronics	**Topologicals**	**Electronics**	**Topologicals**
HOMO	NHOH count	Total charge	Number of H acceptors
LUMO	Fraction of SP3 carbonds	Total energy	Number of H donors
HL-Gap	NO count	Total polarizability alpha	Number of heteroatoms
Fermi Level	Number of aliphatic rings	Total FOD	Number of rotatble bonds
Dipole moment	Number of aromatic rings	TPSA	-
Mol LogP	-	-	-

The screening protocol involved several steps. Firstly, the dataset was curated to remove missing values, correlated descriptors (threshold of 0.9), duplicated values, and variables with very low correlation to the target prediction (noise with a threshold of 0.001). Secondly, models were trained using different parameters and hyperparameters, with the data split into a training set, validation set, and test set. The data were split using k-neighbors clustering-based (KN), with

proportions of 60%, 70%, 80%, and 90%. Thirdly, the models were used to predict the ionization energy of our enumerated subset and our final AI-generated library. From the enumerated subset, it was possible to determine the optimal properties ranges and topological features of molecules with E_i values ranging between 7.5 eV and 8.5 eV.

Goal-Directed Generative Model

The model employed was a generative sequence-based neural network trained to capture the probabilities of generating valid SMILES auto-regressively. (65) In this model, the SMILES are constructed atom by atom from a fixed and known vocabulary (V) of tokens (t), previously defined in an unsupervised trained model using the public database ChEMBL 25, referred to as a prior model. (30,65) Equation 2 (Eq. 2) defines these probabilities, where the joint probability P(T|S) denotes the likelihood of generating a specific sequence T of length l given an input sequence S.

$$P(T|S) = \prod_{i=1}^{l} P(t_i|t_{i-1}, t_{i-2}, t_{i-3}, \ldots, t_i, S) \qquad \text{(Eq. 2)}$$

The Deep Reinforcement Learning (RL) technique was employed, wherein the agent operated within an environment to learn a strategy (policy or goal) by maximizing the reward signal. In molecule generation, the prior model was trained to satisfy a predefined property profile previously ranked with a diverse subset of 500 compounds. (30,87) From the topological and molecular information, it was possible to parameterize the scoring functions of the generative model. The parameters used are listed in Table 2. Additionally, the generative model underwent screening using various parameter values for the learning rate and epochs. The selected learning rates were 0.0001 and 0.001, while the selected epoch values were 100, 500, and 1000. Further parameters of the generative model are detailed in the supplementary information.

Table 2. Scoring functions used to drive the molecules generation

Property	Weight	Transformation function	Range
Molecular Weight (u)	0.34	Double Sigmoid	100 - 1000
SLogP	1	Reverse sigmoid	0.08 - 11.41
TPSA	1	Sigmoid	0.0 - 104.31
Number of Rotatable Bonds	1	Reverse Sigmoid	0.0 - 9.0
Number of HB acceptors	1	Double Sigmoid	0.0 - 9.0
Number of HB donors	1	Reverse Sigmoid	0.0 - 3.0
Fraction of SP3 carbon atoms	0.5	Reverse Sigmoid	0.0 - 0.36
Number of hetero atoms	1	Reverse Sigmoid	0.0 - 9.0

Number of aromatic rings	1	Reverse Sigmoid	1.0 - 7.0
Number of aliphatic rings	1	Reverse Sigmoid	0.0 - 2.0
Unwanted SMARTS	0.79	-	-
SA score	1	-	-

The inception, diversity filter, and randomized SMILES strings were integrated into the generative model to enhance performance. (30,36) Inception plays a pivotal role by guiding the generation process to adhere to established “good” SMILES representations within the desired topological space. (30,65) The diversity filter ensures the creation of a diverse library, each with unique SMILES representations, thereby facilitating a more comprehensive exploration of chemical space. Furthermore, including randomized SMILES enables the random shuffling of atoms within SMILES strings, a technique known to enhance the effectiveness of generative models. (36) Finally, to regulate the influence of score functions on the general likelihood of the sequence, our approach employs the strategy known as “Difference between Augmented and Posterior” (DAP) with a σ value of 128. This technique minimizes the squared loss between the augmented and posterior log-likelihoods described previously, thereby sharpening the agent's focus and ensuring a more effective learning process.37,66 It is worth noting that the DAP approach is more robust for rapid learning.

Results and Discussion

Initial ET MALDI database

The Bemis-Murcko fragmentation process of 30 structures in Figure 3 (commercial structures and internally validated cores) yielded 68 building blocks (BB) encoded in SMILES format, comprising 54 frameworks/ring systems and 14 side chains, which were subsequently deduplicated (Fig. 4). After a network correlation of the 54 fragments obtained from the 30 ET MALDI matrices, we observed a hierarchy network spanning 5 levels with 54 nodes and 106 edges, where benzene (coded BB1 in Fig. 4) is the fundamental core, representing 87.1% of the total building blocks identified. Among the building blocks derived from the seven commercial ET MALDI matrices in the set, only BB1, BB4, BB21, BB29, BB39, BB40, and BB45 belonged to this group, while the remaining building blocks correlate with the internally validated cores employed as an initial library for structural enumeration. The primary distinctions among the commercial structures lie in the number of benzene rings and their chemical bonds, including ring fusion and covalent bonds. Heterocycles are only present in the TER and TFF compounds (Fig. 3), featuring 5-membered rings with sulfur (S) atoms.

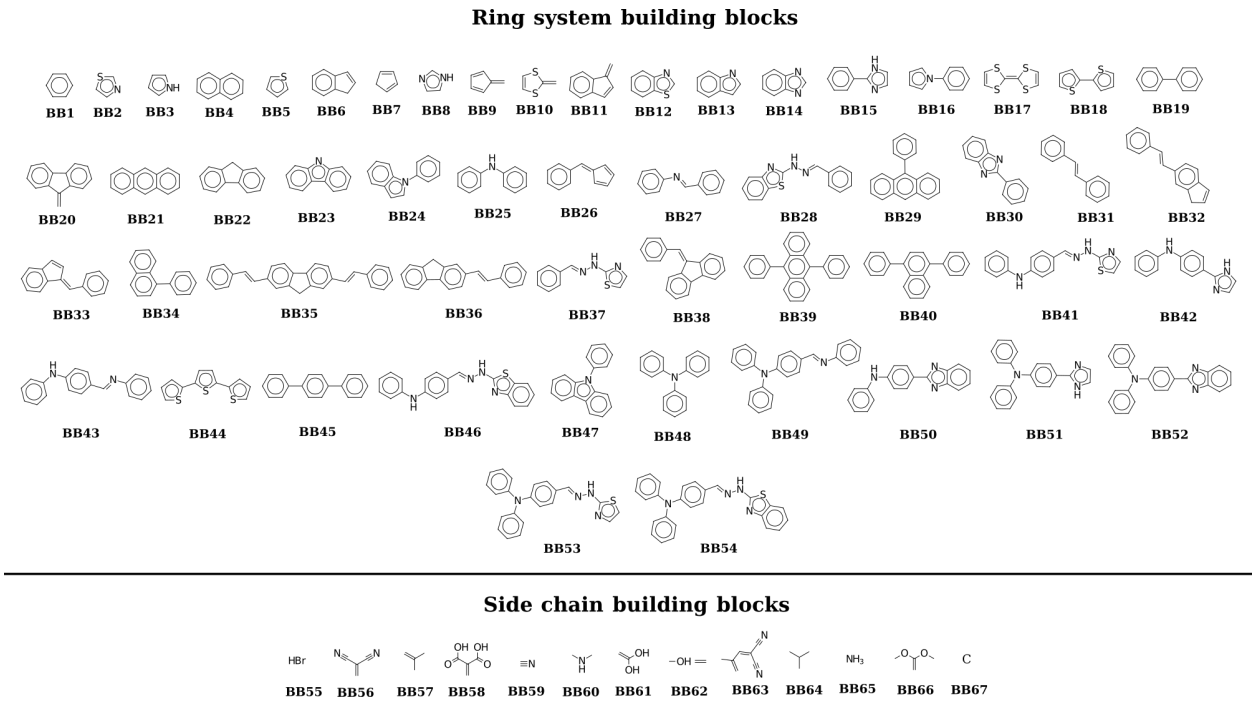


Figure 4. Building blocks from the Bemis-Murcko fragmentation process of the original 30 ET MALDI matrix structures.

The main architectures of the internally tested group comprise triphenylamine, carbazole, and fluorene, with some derivatives containing nitrogen and sulfur atoms, as seen in Fig. 3. Specifically, sulfur-containing derivatives are represented by heterocyclic compounds like N-[(E)-[4-(N-phenylanilino)phenyl]methylideneamino]-1,3-benzothiazol-2-amine (PubChem CID: 9612023), while nitrogen atoms are found in the structure backbone in cyclic and linear forms. Additionally, nitrogen can be present in side chains, such as in the form of a cyano group (-CN), in other proposed structures. The triphenylamine core is an important scaffold of diversification in the generated building blocks, followed by the fluorene and carbazole cores. The extracted side chains as building blocks, mainly derived from the internally proposed structures, comprise electron-withdrawing (EWG) and electron-donating groups (EDG). These side chains affect the physical-chemical properties of potential structures such as UV-vis absorbance, solubility, and E_i, among others.

In the field of MALDI matrix design, Jaskolla and colleagues pioneered the concept of rational design, employing an iterative approach to vary electron-withdrawing (EWG) and electron-donating (EDG) functional groups on an α-cyanocinnamic acid (CCA) core. (22) Subsequent studies on the rational design of novel MALDI matrices follow a similar strategy, focusing on exploring compound families and utilizing chemically conjugated cores with strong UV-vis absorption properties, along with desirable electrical and optical features, and the strategic placement of EWG and EDG groups. (13,14,16-21) The impact of functional group positions within a particular structure on the performance of a compound as a MALDI matrix

was illustrated by Tammekivi *et al.* in their study on the variation of amino positions in monoaminoacridines. (14) Our group has previously used rational design incorporating EWG and EDG functional groups in phenylenevinylene (PV) cores to increase structural complexity and improve E_i, solubility, vacuum stability, desorption performance, and disrupt compound planarity. (24-26) We observed that the presence of key functional groups, such as methoxy (-OCH3), carboxyl (-COOH), hydroxy (-OH), cyano (-CN), and bromine (Br), modulates the core performance as an ET MALDI matrix. (24,26,49) In various other chemical domains, researchers have utilized electron-withdrawing (EWG) and electron-donating (EDG) groups to modify the planarity of compounds and evaluate their effects on photophysical properties, including quantum yield and UV-vis absorbance. (89-91) We employ a similar approach to investigate the chemical and structural implications of particular side-chain building blocks.

Enumerated and AI-generated Libraries

Enumerated library

Enumerating the initial structures from the 68 building blocks, pairing two fragments at a time, generated a library of 82,753 compounds organized into 10 scaffold levels, totaling 103,940 nodes and 186,520 edges. These numbers highlight the remarkably high scaffold diversity within the dataset compared to the initial library. Through clustering, 4,787 clusters were formed comprising 1,117 molecules with only 1.34% being singletons. The Tanimoto average for intracluster similarity was 0.96, and the intercluster index average was 0.37, indicating a highly successful clustering process and comprehensive, diverse sampling of the structural library. Characterizing the enumerated dataset through the CSR curve revealed a p50 index of 0.46 and an area under curve (AUC) of 0.53. In contrast, the initial ET-MALD enumerated subset gave 15 of the initial ET MALDI compounds, including both commercial and internally validated groups. The commercial ET MALDI compound identified was 9,10-DP-ANT (Fig. 3), while the proposed compounds included derivatives based on the carbazole and triphenylamine cores. Fluorene derivatives were absent in our selected subset. Furthermore, the selected subset contained 55 compounds reported in the PubChem database, representing approximately 11% of the total dataset. We observed several interesting trends in the enumerated library associated with topological property distributions. As depicted in Figure 5, properties such as molecular weight (MW), SlogI dataset exhibited a p50 index of 0.18 and an AUC of 0.73. These metrics indicate that in the initial library, only 18% of the scaffolds are needed to generate 50% of the dataset. In other words, half of the 29 compounds can be derived from just 11 scaffolds. However, for the enumerated dataset, generating 50% of the library requires 47,812 scaffolds. This highlights a linear correlation between the number of scaffolds and the number of final generated compounds in our enumerated set, reflecting an ideal relationship for achieving large structural diversity. (92) In contrast, the initial ET MALDI dataset shows that a small number of scaffolds can generate a significant portion of the total library, indicating low structural diversity. Additionally, in terms

of Scaled Shannon Entropy (SSE), the initial ET MALDI dataset exhibited a value of 0.64 (equivalent to 3.2 bits), whereas the enumerated dataset had an SSE value of 0.95 (equivalent to 15.8 bits).

Analysis of ourP, and the number of aromatic rings exhibited relatively homogeneous distribution values across the dataset.

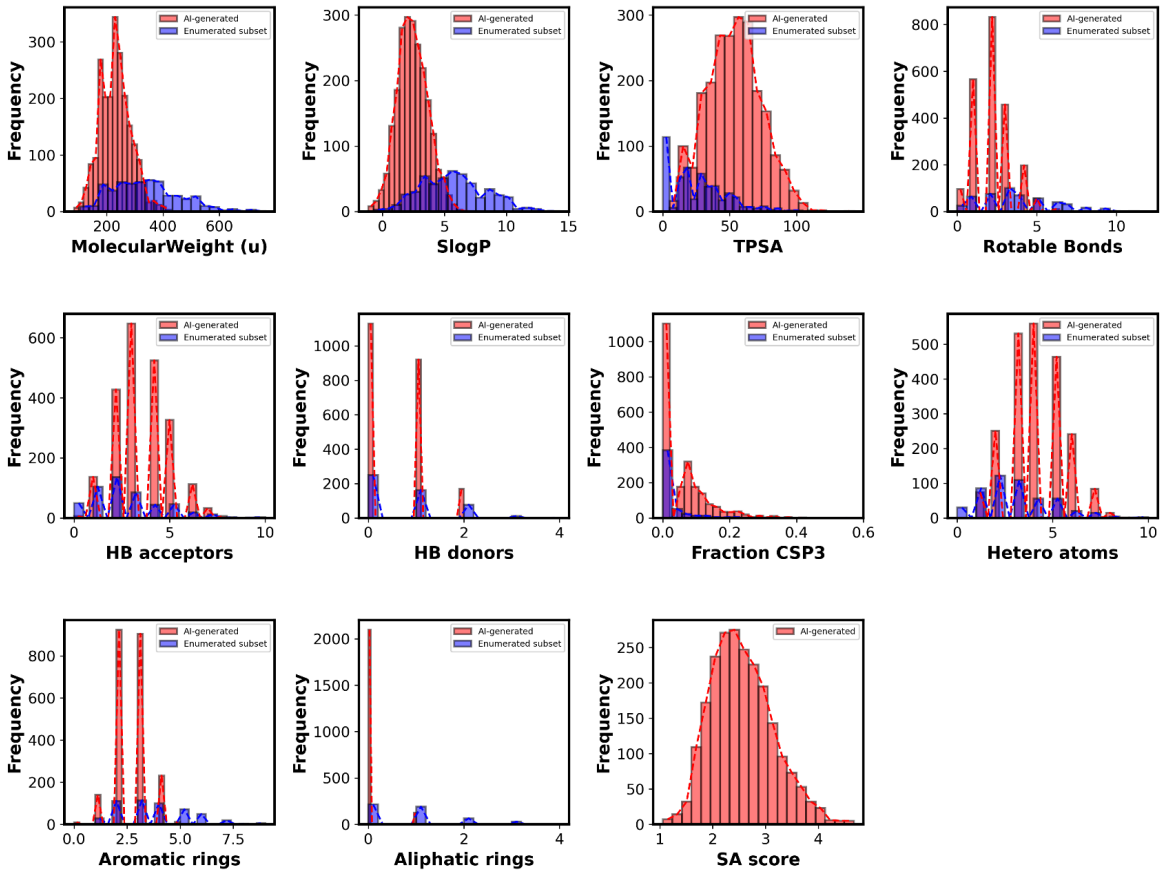


Figure 5. Histograms of properties distribution for enumerated (Blue) and the AI-generated library (Red) within the constraints of Table 2. Only for the AI-generated library, a synthetic accessibility (SA) score was included. (72)

However, other properties displayed asymmetric distributions, suggesting that our enumeration process tends to generate molecules within specific ranges of values. For instance, properties such as TPSA, the number of rotable bonds, HB acceptors and donors, fraction of sp3 carbons, number of heteroatoms (non-heavy), and aliphatic rings tended to have distributions skewed towards lower values. These values will be further discussed in the next section.

Machine Learning Prediction of E_i

Considering the reported and calculated ionization energies of the 30 initial set of ET MALDI matrices, we observed a distribution ranging between 6.5 eV and 8.5 eV (Fig. 6). Based on these values, we curated the data between the E_i and the calculated descriptors listed in Table 1 for our enumerated subset. After data curation, only 17 descriptors were retained, while others were removed due to their low correlation with the target E_i range and high correlation among themselves (> 0.9). (86,93) Specifically, the correlation between total energy and electronic energy descriptors was 1.0, while that of the total dispersion C6 descriptor was 0.93. Conversely, the total charge descriptor showed no correlation with the E_i value range, as the total charge remained constant at zero for each data point. The Fermi-level energy descriptor demonstrated a correlation of 0.92 with the HOMO energy. Moreover, the total polarizability alpha descriptor showed a correlation value of 0.95 with the total dispersion C8 descriptor. Lastly, the fraction CSP3 descriptor correlated 0.0 with the E_i, falling below the threshold of 0.001.

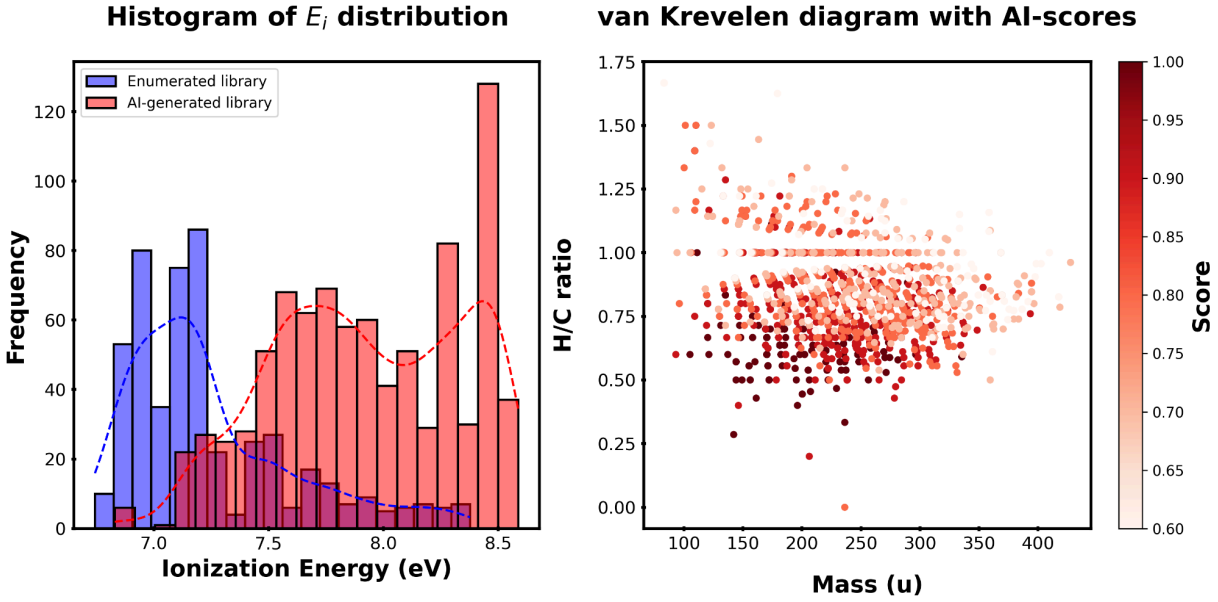


Figure 6. Ionization energy distributions of the initial enumerated (blue) and AI-generated (red) libraries. van Krevelen diagram with AI scores representing the H/C ratio against the mass of the AI-generated final compounds.

The machine learning analysis determined that the Random Forest model showed the highest efficacy in predicting E_i values, using a random state of 233 and a training-validation split of 70:30 as seen in Fig. 7. This finding aligns with previous studies highlighting the robust performance of RF models, particularly with smaller datasets. (94,95) Conversely, the Neural Network (NN) model demonstrated comparatively inferior performance under the same training-validation configuration. While NN models may achieve similar accuracies with larger

and more diverse datasets, RF outperformed NN in this specific context, emphasizing its suitability for the task. As our training process was supervised, we conducted a regression analysis between the calculated or reported E_i and the corresponding predicted values (See Fig. 6). Throughout the training phase, the RF model exhibited commendable performance, achieving an R^2 value of 0.92 with an RMSE of 0.16. Subsequently, during the validation stage, the model maintained strong predictive capability, yielding an R^2 value of 0.88 with a RMSE of 0.14. These metrics collectively indicate the RF model's proficiency in accurately predicting E_i values. Leveraging this trained RF model, we predicted the E_i values of the enumerated subset and the final AI-generated library (Fig. 5).

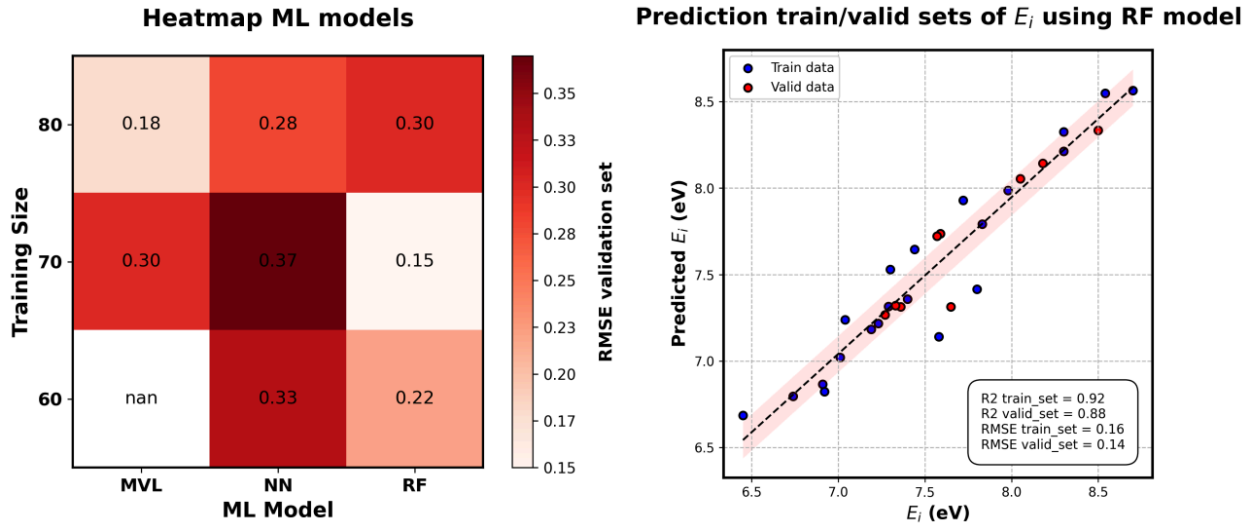


Figure 7. Heatmap showing correlations between the RMSE values and the model utilized for E_i prediction. Regression plot for known vs predicted E_i values calculated using the optimal performance machine learning (ML) model found using ROBERT protocol. (86)

Analysis of the E_i values for both the initial ET MALDI matrices group and the enumerated subset reveals a clear trend: the E_i energies within the enumerated library predominantly cluster toward lower values, typically ranging between 6.74 and 7.40 eV (Figure 6). Structural examination of the scaffolds and side chains in the enumerated subset highlights a notable decrease in the number of side chains and their arrangements compared to the AI-generated library (Fig. 8).

The primary simple side chains identified include -CN, -COOH, Br, -OCH_3, -OH, secondary amine (R-NH-R), tertiary amine (N-R_3), and -CH_3. Side chains containing -CN and -COOH functional groups tend to correspond to molecules with ionization energies exceeding 8.0 eV, while the presence of Br, -OCH_3, -OH, RNHR, and CH_3 functional groups is associated with molecules exhibiting ionization energies under 8.0 eV. Specifically, molecules containing secondary amine and methoxy functional groups display the lowest average E_i values of approximately 7.1 eV. Conversely, more intricate side chains such as

2-(2-methylprop-2-enylidene)propanedinitrile, vinylidene cyanide, and 2-methylene malonic acid correlate with molecules possessing the highest ionization energies in the enumerated subset, with vinylidene cyanide showing the highest E_i value. Additional side chains such as ethene-1,1-diol, ethene, 1,1-dimethoxy-, and 2-methyl-1-butene are also linked to molecules with low E_i values, around 7 eV (Fig. 8).

Analysis of the resulting Murcko scaffolds reveals that the SS30 structure (Fig. 8) shares some structural similarities with the ET MALDI matrices of the type oligo p-phenylenevinylene, previously reported by our group. (24) Furthermore, SS30 also resembles DBDA (PubChem CID: 82590), a MALDI matrix for LMWC analysis, as reported by Ling et al. (18) Among other chemical architectures identified within the enumerated subset, cores such as pyrrole, benzimidazole, 2- phenylimidazole, 1-phenylpyrrole, trans-stilbene, and indole are primarily associated with E_i values exceeding 8.0 eV. It is important to highlight that trans-stilbene is a structural unit of the CNPVs reference molecules. The side chains linked to these cores significantly influence ionization energy values. Conversely, cores such as SS30, SS83, and benzothiazole are predominantly associated with molecules exhibiting E_i values under 8.0 eV. Castellanos et al., in their study on oligo p-phenylenevinylene derivatives, similarly reported E_i values under 8.0 eV, consistent with our findings regarding the SS30 core. (24)

Another notable feature of the entire set of scaffolds and side chains within the enumerated subset is the lack of structural planarity. This observation aligns with previous findings, wherein the absence of planarity, combined with the presence of EWG groups, contributes to increased E_i values and enhanced desorption capability of ET MALDI matrices, as well as improved UV-vis absorption performance. (24,25,96) The disruption of planarity may correlate with the number of rotable bonds, as depicted in Figure 5, where the number of rotable bonds tends to be lower than 5, suggesting the absence of fused aromatic rings, which increases the likelihood of planarity disruption. Conversely, the partition coefficient (SlogP) demonstrates a homogeneous distribution of values across our enumerated subset (see Fig. 5), indicating compounds with a range of lipophilic and hydrophilic properties. (97) Lower SlogP values indicate a hydrophilic feature and enhanced solubility in common polar solvents such as acetonitrile (ACN) or tetrahydrofuran (THF) used during ET MALDI MS sample preparation. (2,55)

A comprehensive analysis of scaffolds and side chains prevalence across the entire enumerated subset reveals a significant abundance of electron-donating functional groups. This observation aligns with findings reported in prior studies, which consistently link these groups to a decrease in ionization energy values. Consequently, the predominant chemical structures within the enumerated subset exhibit E_i values around 7 eV as seen in Fig. 6.

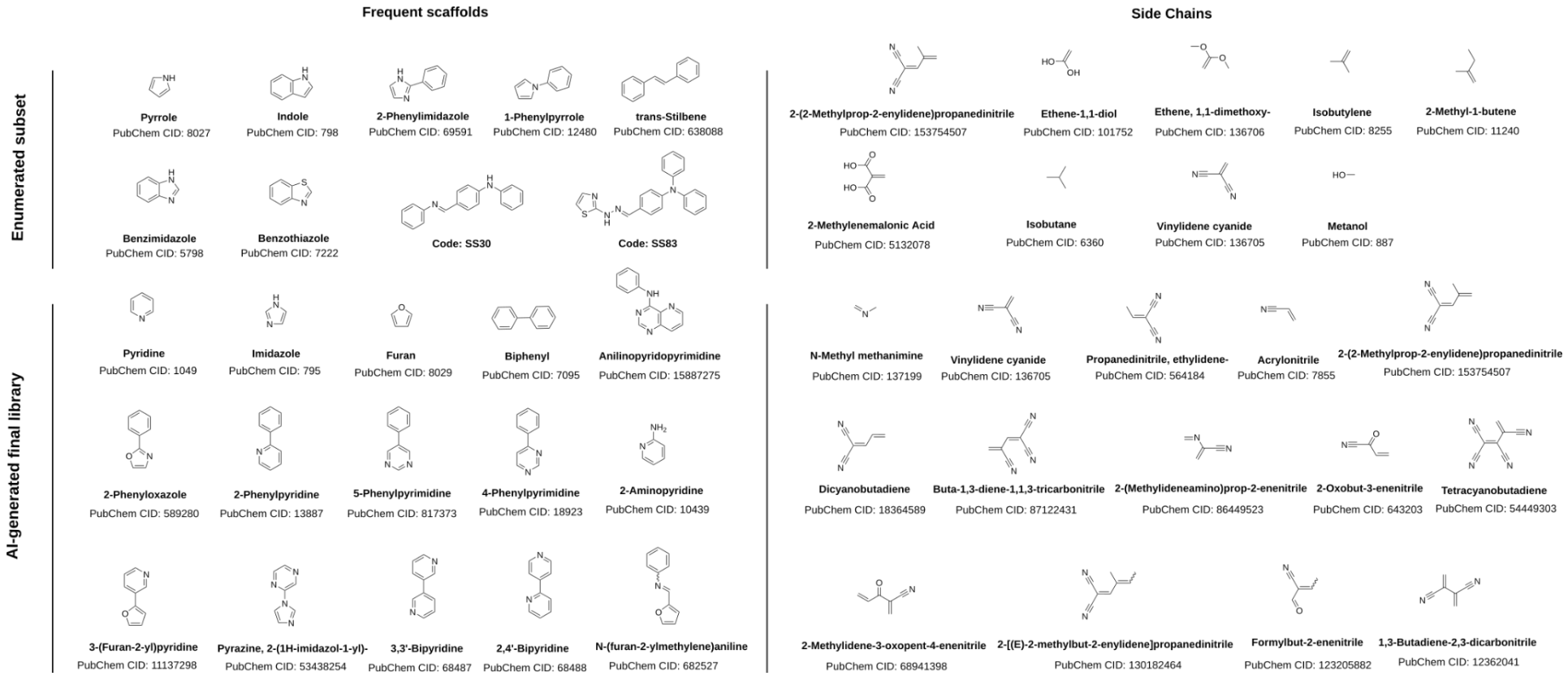


Figure 8. Main scaffolds and side chains from the Enumerated and AI-generated libraries, associated with molecules with E_i between 7.5 eV and 8.5 eV

AI-Generated Library

Given the limited number of structures with high E_i values in our enumerated library comparable to reported E_i for the ET MALDI matrices in our enumerated subset we utilized this knowledge to train a Goal-Directed generative model. This model exclusively employed the descriptors in Table 1, incorporating the values and transformation functions outlined in Table 2. The resulting structures, with E_i values ranging between 7.5 eV and 8.5 eV, got high AI scores, falling within the range of 0.9 to 1.0. Notably, our analysis unveiled two distinct Gaussian distributions at 7.6 eV and 8.4 eV, as illustrated in Figure 6.

While specific ranges of properties and topological features were initially assigned based on our enumerated subset, our generative model occasionally yields structures with altered distributions of property ranges and tendencies, favoring the creation of high AI-scoring and synthetically viable molecules. Consequently, the AI-generated structures exhibit high E_i values. Notable changes in property distributions are observed in parameters such as molecular weight (MW), where lower values around 250 u and under 400 u were favorable (See Figure 5). Similar trends are observed with SlogP, the fraction of sp3 carbons, and the number of aromatic and aliphatic rings. In contrast, properties like TPSA, HB acceptors, and the number of heteroatoms (non-heavy) tend to display increased values compared to the initial distributions of our enumerated subset. Meanwhile, other properties, such as the number of rotable bonds and HB donors, tend to maintain the same distribution as the enumerated subset. The SlogP parameter within the AI-generated library favors the creation of molecules soluble in polar solvents such as MeOH, ACN, and THF.

The AI-generated library highlights the frequent occurrence of scaffolds associated with pyridine, imidazole, furan, biphenyl, 2-phenyloxazole, 2-phenylpyridine, 2-aminopyridine, 5-phenylpyrimidine, and 4-phenylpyrimidine (see Fig. 5). The 4-anilinopyridopyrimidine scaffold is also common, but the E_i values for molecules in this group do not exceed 7.3 eV, regardless of the side chain type. Notably, 386 structures were identified with the highest E_i values exceeding 8.0 eV, among which 206 exhibited ionization energies surpassing 8.3 eV, with 44 structures having 8.5 eV or higher. Remarkably, these molecules attained AI scores exceeding 0.9, indicating alignment with our design objectives. Of the structures with AI scores exceeding 0.9, 291 are already reported in the PubChem database, with an average SA score of 2.6 (indicating ease of synthesis according to the SA score published by Ertl *et al.*). (72) Among these, 116 structures have E_i values ranging from 8.0 eV to 8.6 eV, with repeated scaffolds including pyridine, imidazole, furan, 2-phenylpyridine, and 2-phenyloxazole. Additionally, scaffolds such as N-(furan-2-ylmethylene)aniline, 3-(furan-2- yl)pyridine, 3,3’-bipyridine, pyrazine, 2-(1H-imidazol-1-yl)-, and 2,4’-bipyridine also appear repeatedly with high E_i values.

Regarding side chains, the −CN group is the most frequently occurring, followed by vinylidene cyanide, N-methyl methenamine, propane-dinitrile, ethylidene-, acrylonitrile, and 2-(2-methylprop-2-enylidene)propanedinitrile. However, comparing the enumerated subset with our AI-generated library, the latter exhibits a more diverse range of side chains, predominantly featuring -CN groups and their structural combinations (See Fig. 8). Some representative side chains associated with molecules exhibiting E_i values over 8.4 eV and reported in the PubChem database in the AI-generated final library include dicyanobutadiene, vinylidene cyanide, buta-1,3-diene-1,1, 3-tricarbonitrile, 2-(methylideneamino)prop-2-enenitrile, 2-oxobut-3-enenitrile, tetracyanobutadiene, 2-methylidene-3 -oxopent-4-enenitrile, 2-[(E)- 2-methylbut-2 -enylidene]propanedinitrile, formylbut-2-enenitrile, and 1,3-butadiene-2,3- dicarbonitrile.

Figure 5 shows an average synthesis estimation viability SA score of 2. (94). This score suggests that all the above compounds are potentially easy to synthesize. Scores ranging from 1 to 6 indicate easily obtainable compounds, while scores above 6 indicate difficult-to-synthesize compounds. (70) Our AI-generated library contains compounds that are also present on PubChem. The average SA score of these compounds is 2. (14) This implies that these derivatives can be synthesized and tested as an ET MALDI matrix in the next phase of our research.

The van Krevelen diagram (Fig. 6) provides valuable insights into the unsaturation and the aromaticity of these compounds. Our modified van Krevelen diagram illustrates that the highest AI-scoring compounds predominantly exhibit H/C ratios between 0.50 and 0.75, with molecular weights ranging from 150 u to 300 u. This observation suggests that most compounds generated by our model tend to possess unsaturations, facilitating effective charge distribution and strong absorption in the UV-vis region. On the other hand, the resulting low molecular weights can be associated with the model's preference for two or three aromatic rings only, with few heteroatoms and side chains in the AI-generated compounds. This leads to compounds with low molecular weights like those found in existing commercial and designed molecules. However, the fundamental molecular features of the existing ET MALDI compounds remain intact, albeit with expanded structural diversity. This expansion has resulted in molecules that are structurally different but possess a similar range of properties.

Conclusions

The structural enumeration using Bemis-Murcko fragments derived from a set of commercial and previously designed ET MALDI matrices resulted in a vast and diverse library of compounds that inhabit a similar topological and property space as the initial matrices. This enumerated library was critical in guiding the structural generation process and determining the preferred property ranges for enhanced ET MALDI compounds. Subsequently, this information

enabled the construction of a Goal-Directed generative model, efficiently producing diverse structures within the desired property space. As a result, the generated compounds exhibited higher E_i values and molecular features similar to existing matrices, indicating the possible ideal property ranges for designing efficient ET MALDI compounds. This was evident when comparing our results with the literature, where scaffolds such as trans-stilbene and the SS30 found scaffold closely resembled previously reported structures by Castellanos *et al.*, (24) Pradilla *et al. (*25) and Ling *et al.* (69) Although these structures were not included as training samples for our generative model, it was still capable of generating compounds very similar to them. It was possible to arrive at structures analogous to α-CNPV-CH_3 through the use of artificial intelligence, which were designed purely by empirical approaches.

To potentially generate structures similar to those designed by Pradilla *et al.*, which exhibited excellent performance as MALDI ET matrices, further model refinement may be necessary. This refinement could involve tuning the model with additional specific parameters and scoring functions, such as UV vis absorbance, solubility, quantum yield, and vacuum stability, which are crucial parameters for MALDI matrix design. Additionally, expanding the initial enumerated library beyond the scope of the current study could enable the exploration of a broader topological space, facilitating the generation of structures more closely resembling those of interest.

Future Research Activities

- Continue improving the generative model by incorporating new deep learning and reinforcement learning techniques. This may include the use of more advanced architectures or the integration of additional data to enhance the accuracy of molecular property predictions.
- Synthesize a selection of compounds generated by the AI-generative model for experimental validation. These compounds should be evaluated for their electron transfer (ET) capability in MALDI assays. This step will allow verifying the accuracy of the model's predictions and adjusting the design parameters.
- Make the generative model more robust by expanding the initial database with analogous compounds and more specific physicochemical and optoelectronic properties for the design of ET MALDI matrices, such as UV-vis absorbance, quantum yield, solubility, and electron affinity (for negative ion mode analysis). These more specific properties require greater computational expense and more robust methodologies to parameterize the generative model with customized scoring functions.
- Improve the Ei prediction process by increasing the number of molecules to train the predictive RF model. The molecules can be obtained from databases such as CAS.
- Scale the structural enumeration process to more than one generation using parallel processing configurations with graphics processing units (GPUs) in High Performance Computing (HPC), as well as the training and generation process of the generative model.

Divulgation

Products Related to This Research

1. **Oral presentation**
   - **1.1. Title:** AI-Assisted Design of MALDI matrices for Phytoplankton Chemotaxonomy via *in silico* and *in vitro* Experimentation.
   - 1.2. **Authors:** Carlos A. Padilla, Luis M. Díaz-Sánchez, Cristian Blanco-Tirado, Aldo F. Combariza, and Marianny Y. Combariza.
   - **1.3. Data:** 03-05 May 2023.
   - **1.4. Event:** VIII Encuentro Nacional de Químicos Teóricos y Computacionales.
   - **1.5. Place:** Universidad Nacional de Colombia.
   - **1.6. DOI:** 10.5281/zenodo.10126718
   - 1.7. **Memory publicación:** https://drive.google.com/file/d/1UaJtII_IZvoO5lC_cmTFRg2SeubcwX8S/view?usp=sharing

2. **Oral presentation**
   - 2.1. **Title:** Aventurándose en Nuevos Territorios de Matrices MALDI: Una Red Neuronal Artificial para cartografiar el cosmos químico
   - 2.2. **Authors:** Carlos A. Padilla, Luis M. Díaz-Sánchez, Cristian Blanco-Tirado, Aldo F. Combariza, and Marianny Y. Combariza.
   - 2.3. **Data:** 02-06 October 2023.
   - 2.4. **Event:** XIX Congreso Colombiano de Química
   - 2.5. **Place:** Universidad de los Andes
   - 2.6. **DOI:** 10.5281/zenodo.10126646
   - 2.7. **Memory publicación:** https://s3.amazonaws.com/eventtia/event_files/176954/original/Memorias_141123_%28red%29.pdf?1699981317

3. **Oral presentation**
   - 3.1. **Title:** Exploring the Future of MS: AI-Guided Design of MALDI Matrices using Theoretical and Empirical Insights.
   - 3.2. **Authors:** Carlos A. Padilla, Luis M. Díaz-Sánchez, Cristian Blanco-Tirado, Aldo F. Combariza, and Marianny Y. Combariza.

3.3. **Data:** 02-06 June 2024.
3.4. **Event:** 72nd ASMS Conference on Mass Spectrometry and Allied Topics
3.5. **Place:** Anaheim Convention Center, Anaheim, California.

4. Original Paper

4.1. **Title:** AI-Guided Design of MALDI Matrices: Exploring the Electron Transfer Chemical Space for Mass Spectrometric Analysis of Low Molecular Weight Compounds.
4.2. **Authors:** Carlos A. Padilla, Luis M. Díaz-Sánchez, Cristian Blanco-Tirado, Aldo F. Combariza, and Marianny Y. Combariza.
4.3. **Journal:** Journal of American Society of Mass Spectrometry (JASMS)
4.4. **Status:** Under Review

Other Products During the Master's Program

5. Poster presentation

5.1. **Title:** Comp. Chem. Analysis Of Proton Transference In MALDI For Phytoplankton Chemotaxonomy
5.2. **Authors:** Maria J. Álvarez, Carlos A. Padilla, Luis M. Díaz-Sánchez, Cristian Blanco-Tirado, Aldo F. Combariza, and Marianny Y. Combariza.
5.3. **Data:** 03-05 May 2023.
5.4. **Event:** VIII Encuentro Nacional de Químicos Teóricos y Computacionales.
5.5. **Place:** Universidad Nacional de Colombia.
5.6. **DOI:** 10.5281/zenodo.10076926

6. Poster presentation

6.1. **Title:** Photoinduced Electron Transfer in Donor-Acceptor Systems for MALDI: a DFT Study
6.2. **Authors:** Emmanuel Campo, Carlos A. Padilla, Luis M. Díaz-Sánchez, Cristian Blanco-Tirado, Aldo F. Combariza, and Marianny Y. Combariza.
6.3. **Data:** 03-05 May 2023.
6.4. **Event:** VIII Encuentro Nacional de Químicos Teóricos y Computacionales.
6.5. **Place:** Universidad Nacional de Colombia.
6.6. **DOI:** 10.5281/zenodo.10126585

7. Poster presentation

7.1. **Title:** Quantum Secrets Of Phytoplankton Pigments Analysis By MALDI MS: Proton Affinity Of MALDI Matrices And Phytoplankton Pigments From Computational Chemistry

7.2. **Authors:** Maria J. Álvarez, Carlos A. Padilla, Luis M. Díaz-Sánchez, Cristian Blanco-Tirado, Aldo F. Combariza, and Marianny Y. Combariza.

7.3. **Data:** 04-06 October 2023.

7.4. **Event:** Congreso Internacional de Semilleros de Investigación - Educación -Tecnología Portugal

7.5. **Place:** Portugal

7.6. **DOI:** 10.5281/zenodo.10076991

References

1. Modern Instrumental Analysis 1st ed. (eds Ahuja, S. & Jespersen, N.) 896. isbn: 9780444522597 (2006)

2. de Hoffmann, E.; Stroobant, V. Mass Spectrometry. Principles and Applications, 3rd ed.; Methods in Molecular Biology; Wiley-Interscience, 2007.

3. Awad, H., Khamis, M. M. & El-Aneed, A. Mass Spectrometry, Review of the Basics: Ionization. Applied Spectroscopy Reviews 50, 158–175. issn: 0570-4928. http://www.tandfonline.com/doi/abs/10.1080/05704928.2014. 954046 (Feb. 2015)

4. Eidhammer, I., Flikka, K., Martens, L. & Mikalsen, S.-O. Computational methods for Mass Spectrometry proteomics 1st ed., 285. isbn: 0470512970, 9780470512975, 9780470724293 (John Wiley & Sons, Chichester, England; Hoboken, NJ, 2007)

5. Smoluch, M., Grasso, G., Suder, P. & Silberring, J. Mass Spectrometry: An Applied Approach, 2nd Edition isbn: 978-1-119-37736-8 (2019)

6. Lee, M. S.; Ji, Q. C. Protein Analysis Using Mass Spectrometry; Lee, M. S., Ji, Q. C., Eds.; John Wiley & Sons, Inc.: Hoboken, NJ, USA, 2017. https://doi.org/10.1002/9781119371779

7. Rune, M. Mass Spectrometry Data Analysis in Proteomics isbn: 1-59745-275-0. http://link.springer.com/10.1385/1597452750 (Humana Press, New Jersey, Nov. 2006

8. Evans, C. A., Wright, P. C. & Noirel, J. Mass Spectrometry of Proteins (eds Evans, C. A., Wright, P. C. & Noirel, J.) isbn: 978-1-4939-9231-7. http : //www.springer.com/series/7651%20http://link.springer.com/10. 1007/978-1-4939-9232-4 (Springer New York, New York, NY, 2019)

9. Medhe, S. Ionization Techniques in Mass Spectrometry: A Review. Mass Spectrometry & Purification Techniques 04, 1–6. issn: 24699861. https :/ / www . omicsonline . org / open - access / ionization - techniques - in -mass-spectrometry-a-review-2469-9861-1000126-102758.html (2018).

10. Tanaka, K. et al. Protein and polymer analyses up tom/z 100 000 by laser ionization time-of-flight mass spectrometry. Rapid Communications in Mass Spectrometry 2, 151–153. issn: 0951-4198. http : / / doi . wiley . com / 10 .1002/rcm.1290020802 (Aug. 1988).

11. Karas, M. & Hillenkamp, F. Laser Desorption Ionization of Proteins with Molecular Masses Exceeding 10000 Daltons. Analytical Chemistry 60, 2299–2301. issn: 15206882. https://pubs.acs.org/doi/abs/10.1021/ac00171a028 (Oct. 1988).

12. Vasil'ev, Y. V. et al. Electron Transfer Reactivity in Matrix-Assisted Laser Desorption/Ionization (MALDI): Ionization Energy, Electron Affinity and Performance of the DCTB Matrix within the Thermochemical Framework. The Journal of Physical Chemistry A 110, 5967–5972. issn: 1089-5639. https://pubs.acs.org/doi/10.1021/jp060568f (May 2006)

13. Chen, R.; Chen, S.; Xiong, C.; Ding, X.; Wu, C.-C.; Chang, H.-C.; Xiong, S.; Nie, Z. N-(1-Naphthyl) Ethylenediamine Dinitrate: A New Matrix for Negative Ion MALDI-TOF MS Analysis of Small Molecules. J. Am. Soc. Mass Spectrom. 2012, 23 (9), 1454–1460. https://doi.org/10.1007/s13361-012-0421-z.

14. Tammekivi, E.; Ghiami-Shomami, A.; Tshepelevitsh, S.; Trummal, A.; Ilisson, M.; Selberg, S.; Vahur, S.; Teearu, A.; Lõkov, M.; Peets, P.; Pagano, T.; Leito, I. Experimental and Computational Study of Aminoacridines as MALDI(-)-MS Matrix Materials for the Analysis of Complex Samples. J. Am. Soc. Mass Spectrom. 2021, 32 (4), 1080–1095. https://doi.org/10.1021/jasms.1c00037

15. Wiangnon, K.; Cramer, R. Sample Preparation: A Crucial Factor for the Analytical Performance of Rationally Designed Maldi Matrices. Anal. Chem. 2015, 87 (3), 1485–1488. https://doi.org/10.1021/ac504412p.

16. Huang, P.; Huang, C.-Y.; Lin, T.-C.; Lin, L.-E.; Yang, E.; Lee, C.; Hsu, C.-C.; Chou, P.-T. Toward the Rational Design of Universal Dual Polarity Matrix for MALDI Mass Spectrometry. Anal. Chem. 2020, 92 (10), 7139–7145. https://doi.org/10.1021/acs.analchem.0c00570.

17. Yang, Y.; Gao, D.; Qian, R.; Jiang, Y. Polydopamine-Modified TS-1 Zeolite Framework Nanoparticles as a Matrix for the Analysis of Small Molecules by MALDI-TOF MS. ACS Omega 2020, 5 (32), 19952–19959. https://doi.org/10.1021/acsomega.0c00992.

18. Ling, L.; Li, Y.; Wang, S.; Guo, L.; Xiao, C.; Chen, X.; Guo, X. DBDA as a Novel Matrix for the Analyses of Small Molecules and Quantification of Fatty Acids by Negative Ion MALDI-TOF MS. J. Am. Soc. Mass Spectrom. 2018, 29 (4), 704–710. https://doi.org/10.1007/s13361-017-1881-y.

19. He, Q.; Chen, S.; Wang, J.; Hou, J.; Wang, J.; Xiong, S.; Nie, Z. 1-Naphthylhydrazine Hydrochloride: A New Matrix for the Quantification of Glucose and Homogentisic Acid in Real Samples by MALDI-TOF MS. Clin. Chim. Acta 2013, 420, 94–98. https://doi.org/10.1016/j.cca.2012.10.015.

20. He, H.; Qin, L.; Zhang, Y.; Han, M.; Li, J.; Liu, Y.; Qiu, K.; Dai, X.; Li, Y.; Zeng, M.; Guo, H.; Zhou, Y.; Wang, X. 3,4-Dimethoxycinnamic Acid as a Novel Matrix for Enhanced In Situ Detection and Imaging of Low-Molecular-Weight Compounds in Biological Tissues by MALDI-MSI. Anal. Chem. 2019, 91 (4), 2634–2643. https://doi.org/10.1021/acs.analchem.8b03522.

21. Yao, C.; Niu, C.; Na, N.; He, D.; Ouyang, J. Aggregation-Induced Emission Compounds as New Assisted Matrices for Laser Desorption/Ionization Time-of-Flight Mass Spectrometry. Anal. Chim. Acta 2015, 853, 375–383. https://doi.org/10.1016/j.aca.2014.09.047

22. Jaskolla, T. W.; Lehmann, W.-D.; Karas, M. 4-Chloro- -Cyanocinnamic Acid Is an Advanced, Rationally Designed MALDI Matrix. Proc. Natl. Acad. Sci. 2008, 105 (34), 12200–12205. https://doi.org/10.1073/pnas.0803056105.

23. Cristancho, L. Diseñno racional y síntesis de nuevas matrices MALDI de Transferencia Electrónica basadas en 2,7-Dibromofluoreno. PhD thesis (2016).

24. Castellanos-García, L. J.; Agudelo, B. C.; Rosales, H. F.; Cely, M.; Ochoa-Puentes, C.; Blanco-Tirado, C.; Sierra, C. A.; Combariza, M. Y. Oligo P-Phenylenevinylene Derivatives as Electron Transfer Matrices for UV-MALDI. J. Am. Soc. Mass Spectrom. 2017, 28 (12), 2548–2560. https://doi.org/10.1007/s13361-017-1783-z

25. Ramírez-Pradilla, J. S.; Blanco-Tirado, C.; Combariza, M. Y. Electron-Transfer Ionization of Nanoparticles, Polymers, Porphyrins, and Fullerenes Using Synthetically Tunable α-Cyanophenylenevinylenes as UV MALDI-MS Matrices. ACS Appl. Mater. Interfaces 2019a, 11 (11), 10975–10987. https://doi.org/10.1021/acsami.8b22246.

26. Ramírez-Pradilla, J. S.; Blanco-Tirado, C.; Hubert-Roux, M.; Giusti, P.; Afonso, C.; Combariza, M. Y. Comprehensive Petroporphyrin Identification in Crude Oils Using Highly Selective Electron Transfer Reactions in MALDI-FTICR-MS. Energy & Fuels 2019b, 33 (5), 3899–3907. https://doi.org/10.1021/acs.energyfuels.8b04325.

27. Díaz-Sánchez, L. M., Blanco-Tirado, C., & Combariza, M. Y. (2023). Electron-transfer MALDI MS methodology for microalgae/phytoplankton pigments analysis. MethodsX, Vol. 10, p. 102140. doi:10.1016/j.mex.2023.102140

28. Choudhary, N., Bharti, R., & Sharma, R. (2021). Role of artificial intelligence in chemistry. Materials Today: Proceedings, 48, 1527–1533. https://doi.org/10.1016/j.matpr.2021.09.428

29. Schneider, G., Fechner, U. Computer-based de novo design of drug-like molecules. Nat Rev Drug Discov 4, 649–663 (2005). https://doi.org/10.1038/nrd1799

30. Olivecrona, M., Blaschke, T., Engkvist, O. et al. Molecular de-novo design through deep reinforcement learning. J Cheminform 9, 48 (2017). https://doi.org/10.1186/s13321-017-0235-x

31. Gómez-Bombarelli, R., Aguilera-Iparraguirre, J., Hirzel, T. et al. Design of efficient molecular organic light-emitting diodes by a high-throughput virtual screening and experimental approach. Nature Mater 15, 1120–1127 (2016). https://doi.org/10.1038/nmat4717

32. Sousa, T., Correia, J., Pereira, V., & Rocha, M. (2021). Generative Deep Learning for Targeted Compound Design. In Journal of Chemical Information and Modeling (Vol. 61, Issue 11, pp. 5343–5361). American Chemical Society. https://doi.org/10.1021/acs.jcim.0c01496

33. Bilodeau, C., Jin, W., Jaakkola, T., Barzilay, R., & Jensen, K. F. (2022). Generative models for molecular discovery: Recent advances and challenges. In Wiley Interdisciplinary Reviews: Computational Molecular Science (Vol. 12, Issue 5). John Wiley and Sons Inc. https://doi.org/10.1002/wcms.1608

34. Meyers, J., Fabian, B., & Brown, N. (2021). De novo molecular design and generative models. In Drug Discovery Today (Vol. 26, Issue 11, pp. 2707–2715). Elsevier Ltd. https://doi.org/10.1016/j.drudis.2021.05.019

35. Anstine, D. M., & Isayev, O. (2023). Generative Models as an Emerging Paradigm in the Chemical Sciences. In Journal of the American Chemical Society (Vol. 145, Issue 16, pp. 8736–8750). American Chemical Society. https://doi.org/10.1021/jacs.2c13467

36. Arús-Pous, J., Johansson, S. V., Prykhodko, O., Bjerrum, E. J., Tyrchan, C., Reymond, J.-L., Chen, H., & Engkvist, O. (2019). Randomized SMILES strings improve the quality of molecular generative models. Journal of Cheminformatics, 11(1), 71. https://doi.org/10.1186/s13321-019-0393-0

37. Quirós, M., Gražulis, S., Girdzijauskaitė, S. et al. Using SMILES strings for the description of chemical connectivity in the Crystallography Open Database. J Cheminform 10, 23 (2018). https://doi.org/10.1186/s13321-018-0279-6

38. Kwak, H. S., An, Y., Giesen, D. J., Hughes, T. F., Brown, C. T., Leswing, K., Abroshan, H. & Halls, M. D. (2022). Design of Organic Electronic Materials With a Goal-Directed Generative Model Powered by Deep Neural Networks and High-Throughput Molecular Simulations. Frontiers in Chemistry, 9. https://doi.org/10.3389/fchem.2021.800370

39. Saldívar-González, F.I., Huerta-García, C.S. & Medina-Franco, J.L. Chemoinformatics-based enumeration of chemical libraries: a tutorial. J Cheminform 12, 64 (2020). https://doi.org/10.1186/s13321-020-00466-z

40. Voršilák, M., Kolář, M., Čmelo, I., & Svozil, D. (12 2020). SYBA: Bayesian estimation of synthetic accessibility of organic compounds. Journal of Cheminformatics, Vol. 12, p. 35. doi:10.1186/s13321-020-00439-2

41. Leopold, J.; Popkova, Y.; Engel, K. M.; Schiller, J. Recent Developments of Useful MALDI Matrices for the Mass Spectrometric Characterization of Lipids. Biomolecules 2018, 8 (4). https://doi.org/10.3390/biom8040173.

42. Giraldo-Dávila, D.; Chacón-Patiño, M. L.; Ramirez-Pradilla, J. S.; Blanco-Tirado, C.; Combariza, M. Y. Selective Ionization by Electron-Transfer MALDI-MS of Vanadyl Porphyrins from Crude Oils. Fuel 2018, 226 (March), 103–111. https://doi.org/10.1016/j.fuel.2018.04.016.

43. Mathew D. Halls, Daisuke Yoshidome, Thomas J. L. Mustard, Alexander Goldberg, H. Shaun Kwak & Jacob Gavartin. (2015). Atomic-scale Simulation for the Analysis, Optimization and Accelerated Development of Organic Optoelectronic Materials. Journal of the Imaging Society of Japan, 54(6), 561-569. https://doi.org/10.11370/isj.54.561

44. Janet, J. P., Liu, F., Nandy, A., Duan, C., Yang, T., Lin, S. & Kulik, H. J. (2019). Designing in the Face of Uncertainty: Exploiting Electronic Structure and Machine Learning Models for Discovery in Inorganic Chemistry. Inorganic Chemistry, 58(16), 10592-10606. https://doi.org/10.1021/acs.inorgchem.9b00109

45. Matsuzawa, N. N., Arai, H., Sasago, M., Fujii, E., Goldberg, A., Mustard, T. J., … Halls, M. D. (2020). Massive Theoretical Screen of Hole Conducting Organic Materials in the Heteroacene Family by Using a Cloud Computing Environment. The Journal of Physical Chemistry A. doi:10.1021/acs.jpca.9b10998

46. Suzuki, T.; Midonoya, H.; Shioi, Y. Analysis of chlorophylls and their derivatives by matrix-assisted laser desorption/ionization–time-of-flight mass spectrometry. Analytical Biochemistry 2009, 390, 57–62.

47. Calvano, C. D.; Ventura, G.; Trotta, M.; Bianco, G.; Cataldi, T. R.; Palmisano, F. Electron-Transfer secondary reaction matrices for MALDI MS analysis of Bacteriochlorophyll a in Rhodobacter sphaeroides and its zinc and copper analogue pigments. Journal of The American Society for Mass Spectrometry 2016, 28, 125–135.

48. Calvano, C. D.; Ventura, G.; Cataldi, T. R.; Palmisano, F. Improvement of chlorophyll identification in foodstuffs by MALDI ToF/ToF mass spectrometry using 1, 5-diaminonaphthalene electron transfer secondary reaction matrix. Analytical and bioanalytical chemistry 2015, 407, 6369–6379.

49. Padilla, C. A.; Díaz-Sánchez, L. M.; Combariza, M. Y.; Blanco-Tirado, C.; Combariza, A. F. Photon Harvesting Molecules: Ionization Potential from Quantum Chemical

Calculations of Phytoplanktonic Pigments for MALDI-MS Analysis. Orinoquia 2021, 25, 13–23.

50. Persson, S.; S¨onksen, C. P.; Frigaard, N.-U.; Cox, R. P.; Roepstorff, P.; Miller, M. Pigments and proteins in green bacterial chlorosomes studied by matrix-assisted laser desorption ionization mass spectrometry. European Journal of Biochemistry 2000, 267, 450–456.

51. Gatlin, C. L.; White, K. Y.; Tracy, M. B.; Wilkins, C. E.; Semmes, O. J.; Nyal- widhe, J. O.; Drake, R. R.; Malyarenko, D. I. Enhancement in MALDI-TOF MS anal- ysis of the low molecular weight human serum proteome. Journal of Mass Spectrometry 2011, 46, 85–89.

52. Calderaro, A.; Chezzi, C. MALDI-TOF MS: A Reliable Tool in the Real Life of the Clinical Microbiology Laboratory. Microorganisms 2024, 12.

53. Calvano, C. D.; Monopoli, A.; Cataldi, T. R.; Palmisano, F. MALDI matrices for low molecular weight compounds: an endless story? 2018.

54. Qiao, Z.; Lissel, F. MALDI Matrices for the Analysis of Low Molecular Weight Compounds: Rational Design, Challenges and Perspectives. 2021.

55. Gross, J. H. Mass Spectrometry; Springer International Publishing, 2017.

56. McCarley, T.D., McCarley, R.L., Limbach, P.A.: Electron-transfer ioni- zation in matrix-assisted laser desorption/ionization mass spectrometry. Anal. Chem. 70, 4376–4379 (1998).

57. Knochenmuss, R., A quantitative model of ultraviolet matrix-assisted laser desorption/ionization. J. Mass Spectrom. 2002, 37 (8), 867-877.

58. Knochenmuss, R.; Zenobi, R., MALDI Ionization: The Role of In-Plume Processes. Chem. Rev. 2003, 103 (2), 441-452.

59. George, M.; Wellemans, J. M.; Cerny, R. L.; Gross, M. L.; Li, K.; Cavalieri, E. L. Matrix design for matrix-assisted laser desorption ionization: Sensitive determination of PAH-DNA adducts. Journal of the American Society for Mass Spectrometry 1994, 5, 1021–1025.

60. Korshunova, M.; Huang, N.; Capuzzi, S.; Radchenko, D. S.; Savych, O.; Moroz, Y. S.; Wells, C. I.; Willson, T. M.; Tropsha, A.; Isayev, O. Generative and reinforcement learning approaches for the automated de novo design of bioactive compounds. Communications Chemistry 2022, 5.

61. Baum, Z. J.; Yu, X.; Ayala, P. Y.; Zhao, Y.; Watkins, S. P.; Zhou, Q. Artificial Intelligence in Chemistry: Current Trends and Future Directions. 2021.

62. Zheng, P.; Zubatyuk, R.; Wu, W.; Isayev, O.; Dral, P. O. Artificial intelligence-enhanced quantum chemical method with broad applicability. Nature Communications 2021, 12.

63. Shree, S. S.; Law, J. N.; Tripp, C. E.; Duplyakin, D.; Skordilis, E.; Biagioni, D.; Paton, R. S.; John, P. C. S. Multi-objective goal-directed optimization of de novo stable organic radicals for aqueous redox flow batteries. Nature Machine Intelligence 2022.

64. Hachmann, J.; Afzal, M. A. F.; Haghighatlari, M.; Pal, Y. Building and deploying a cyberinfrastructure for the data-driven design of chemical systems and the exploration of chemical space. Molecular Simulation 2018, 44, 921–929.

65. Loeffler, H. H.; He, J.; Tibo, A.; Janet, J. P.; Voronov, A.; Mervin, L.; Engkvist, O. REINVENT 4: Modern AI-Driven Generative Molecule Design.

66. Ivanov, N. N.; Shulga, D. A.; Palyulin, V. A. Decomposition of Small Molecules for Fragment-Based Drug Design. Biophysica 2023, 3, 362–372.

67. Liu, T.; Naderi, M.; Alvin, C.; Mukhopadhyay, S.; Brylinski, M. Break Down in Order to Build Up: Decomposing Small Molecules for Fragment-Based Drug Design with eMolFrag. Journal of Chemical Information and Modeling 2017, 57, 627–631.

68. Cheng, A. H.; Cai, A.; Miret, S.; Malkomes, G.; Phielipp, M.; Aspuru-Guzik, A. Group SELFIES: a robust fragment-based molecular string representation. Digital Discovery 2023, 2, 748–758.

69. Lim, J.; Hwang, S. Y.; Moon, S.; Kim, S.; Kim, W. Y. Scaffold-based molecular design with a graph generative model. Chemical Science 2020, 11, 1153–1164.

70. Alegre-Requena, J. V.; V., S. S. S.; P´erez-Soto, R.; Alturaifi, T. M.; Paton, R. S. AQME: Automated quantum mechanical environments for researchers and educators. Wiley Interdisciplinary Reviews: Computational Molecular Science 2023, 13.

71. Roy, S.; Llewellyn, C.; Egeland, E.; Johnsen, G. Phytoplankton Pigments: Characterization, Chemotaxonomy and Applications in Oceanography; Cambridge Environmental Chemistry Series; Cambridge University Press, 2011.

72. Ertl, P.; Schuffenhauer, A. Estimation of synthetic accessibility score of drug-like molecules based on molecular complexity and fragment contributions. Journal of Cheminformatics 2009, 1.

73. Bemis, G. W.; Murcko, M. A. The Properties of Known Drugs. 1. Molecular Frameworks. Journal of Medicinal Chemistry 1996, 39, 2887–2893.

74. Scott, O. B.; Chan, A. W. E. ScaffoldGraph: an open-source library for the generation and analysis of molecular scaffold networks and scaffold trees. Bioinformatics 2020, 0–0.

75. G., V.; JA, D.; M., H.; MAF, H. J. A. ChemLG – A program suite for the generation of compound libraries and the survey of chemical space. 2019; https://github.com/hachmannlab/chemlg.

76. Medina-Franco, J. L.; Mart´ınez-Mayorga, K.; Bender, A.; Scior, T. Scaffold diversity analysis of compound data sets using an entropy-based measure. QSAR and Combinatorial Science 2009, 28, 1551–1560.

77. Shannon, C. E. A Mathematical Theory of Communication. The Bell System Technical Journal 1948, 27, 379–423.

78. Vivek-Ananth, R.; Sahoo, A. K.; Baskaran, S. P.; Samal, A. Scaffold and Structural Diversity of the Secondary Metabolite Space of Medicinal Fungi. ACS Omega 2023, 8, 3102–3113.

79. Butina, D. Unsupervised database clustering based on daylight's fingerprint and Tanimoto similarity: A fast and automated way to cluster small and large datasets. Journal of Chemical Information and Computer Sciences 1999, 39, 747–750.

80. Jarvis, R.; Patrick, E. Clustering Using a Similarity Measure Based on Shared Near Neighbors. IEEE Transactions on Computers 1973, C-22, 1025–1034.

81. Gillet, V. J. Diversity selection algorithms. 2011.

82. Smith, J. S.; Zubatyuk, R.; Nebgen, B.; Lubbers, N.; Barros, K.; Roitberg, A. E.; Isayev, O.; Tretiak, S. The ANI-1ccx and ANI-1x data sets, coupled-cluster and density functional theory properties for molecules. Scientific Data 2020, 7.

83. Koopmans, T.U¨ ber die Zuordnung von Wellenfunktionen und Eigenwerten zu den Einzelnen Elektronen Eines Atoms. Physica 1934, 1, 104–113.

84. Jensen, F. Introduction to Computational Chemistry; Wiley, 2007.

85. K., R.; G., D.; K., N. Computational Chemistry and Molecular Modeling; Springer Berlin Heidelberg, 2008.

86. Dalmau, D.; Alegre-Requena, J. V. ROBERT: Bridging the Gap between Machine Learning and Chemistry. ChemRxiv 2023.

87. Blaschke, T.; Aru´s-Pous, J.; Chen, H.; Margreitter, C.; Tyrchan, C.; Engkvist, O.; Papadopoulos, K.; Patronov, A. REINVENT 2.0: An AI Tool for de Novo Drug Design. Journal of Chemical Information and Modeling 2020, 60, 5918–5922.

88. Fialkov´a, V.; Zhao, J.; Papadopoulos, K.; Engkvist, O.; Bjerrum, E. J.; Kogej, T.; Patronov, A. LibINVENT: Reaction-based Generative Scaffold Decoration for in Silico Library Design. Journal of Chemical Information and Modeling 2021.

89. Musil, Z.; Zimcik, P.; Miletin, M.; Kopecky, K.; Petrik, P.; Lenco, J. Influence of electron-withdrawing and electron-donating substituents on photophysical properties of azaphthalocyanines. Journal of Photochemistry and Photobiology A: Chemistry 2007, 186, 316–322.

90. Zhang, L.; Zeng, Z.; Wu, S.; Luo, T.; Li, Z.; Zhang, W.; Wu, W.; Liu, H.; Liu, F. Comparative study on twelve kinds of electron donors for organic functional materials. Dyes and Pigments 2023, 217, 111410.

91. Song, X.-F.; Jiang, C.; Li, N.; Miao, J.; Li, K.; Yang, C. Simultaneously enhancing the planarity and electron-donating capability of donors for through-space charge transfer TADF towards deep-red emission. Chemical Science 2023, 14, 12246–12254.

92. Langdon, S. R.; Brown, N.; Blagg, J. Scaffold diversity of exemplified medicinal chemistry space. Journal of Chemical Information and Modeling 2011, 51, 2174–2185.

93. Artrith, N.; Butler, K. T.; Coudert, F. X.; Han, S.; Isayev, O.; Jain, A.; Walsh, A. Best practices in machine learning for chemistry. 2021.

94. Breiman, L. Random Forests. Machine Learning 2001, 45, 5–32.

95. Ahmad, M. W.; Mourshed, M.; Rezgui, Y. Trees vs Neurons: Comparison between random forest and ANN for high-resolution prediction of building energy consumption. Energy and Buildings 2017, 147, 77–89.

96. Pond, S. J. K.; Rumi, M.; Levin, M. D.; Parker, T. C.; Beljonne, D.; Day, M. W.; Br´edas, J.-L.; Marder, S. R.; Perry, J. W. One- and Two-Photon Spectroscopy of Donor Acceptor Donor Distyrylbenzene Derivatives: Effect of Cyano Substitution and Distor- tion from Planarity. The Journal of Physical Chemistry A 2002, 106, 11470–11480.

97. Pyka, A.; Babu´ska, M.; Zachariasz, M. A comparison of theoretical methods of calculation of partition coefficients for selected drugs. Acta poloniae pharmaceutica 2006, 63, 159–67.